



**University of
Nottingham**

UK | CHINA | MALAYSIA

Interpretability and Annotation Scarcity in Deep Medical Image Segmentation

Thesis submitted to the University of Nottingham for the degree of
Doctor of Philosophy, March 2024.

Golnar Khalili Zadeh Mahani

14342206

Supervised by

Prof. Andrew French

Dr Xin Chen

Prof. Nikolaos Evangelou

Prof. Stamatios Sotiropoulos

Signature _____

Date ____ / ____ / ____

Abstract

This thesis presents an exploration into the challenges of medical image segmentation, particularly focusing on reliability, interpretability, and strong annotation sparsity, through the use of tailored loss functions. It tackles three primary issues: 1) enhancing segmentation output quality 2) addressing the scarcity of strong annotations, and 3) evaluating novel application of the techniques on Multiple Sclerosis (MS) lesion segmentation.

First, this thesis proposes the integration of uncertainty maps into medical image segmentation to gain insights into the confidence of a model. The accuracy of these maps in conveying true uncertainty is a subject of debate. This thesis tackles the intricacies of interpreting uncertainty maps in medical image segmentation, considering factors like lesion shape, border irregularity, inconsistencies in acquisition methods, variance in lesion texture, and the presence of image artifacts. The experiments explore the interpretation of these maps and evaluate their reliability in different domains. The proposed tailored loss functions incorporate model uncertainty, error feedback, and their combinations with uncertainty inherent from the data as a regularisation term.

Second, the thesis addresses the challenge of limited accurate annotations by introducing a tailored loss function designed for weakly-supervised medical image segmentation that relies solely

on bounding box annotations. This novel loss function leverages predictive uncertainty estimates during training to direct model learning towards areas of the image with high confidence. It also integrates a local spatial constraint within the loss function to ensure consistent predictions within a given area. The proposed approach outperforms both state-of-the-art learning-based and non-learning-based methods, showcasing its superior efficacy.

Third, **the thesis examines the use of proposed uncertainty-aware loss functions in the fully supervised segmentation of MS brain lesions.** Despite progress, the challenge of accurately segmenting MS lesions and minimising misdiagnosis remains, emphasising the necessity for improved imaging biomarkers and robust segmentation techniques to increase diagnostic accuracy and patient care in MS. Experiments evaluating two preprocessing methods for extracting brain tissue demonstrate the influence of preprocessing choices on the outcomes of deep segmentation methods in complex datasets. Furthermore, we find that loss functions incorporating model uncertainty feedback using data uncertainty as a regularisation term enhance segmentation performance on unseen MS data acquired from different sites and scanners. This highlights the importance of developing customised loss functions to improve model robustness, in MS lesion segmentation models.

Acknowledgements

In recognition of his dedication and the pivotal role he has played in my PhD journey, I would like to extend my deepest gratitude to my supervisor, Prof. Andrew French, whose mentorship has been instrumental not only in my academic and professional development but also in my personal growth. His patience and positivity have always motivated me, helping me face challenges with hope and resilience. His influence has truly encouraged me to pursue my goals with determination and integrity.

I am deeply grateful to my entire supervision team for their invaluable contribution throughout my academic journey. Their wisdom, and guidance, have been fundamental to my success and have profoundly enriched my experience.

I would also like to thank my family members, whose support has been a constant and comforting presence, even from thousands of miles away. Despite the geographical distances that separate us, their love have bridged the gap, reminding me of the strength and support that comes from home.

I dedicate my Thesis to my loving partner, for it was for his unwavering support, endless patience, and love that carried me through the toughest of times.

Contents

Abstract	i
Acknowledgements	iii
List of Tables	x
List of Figures	xi
Abbreviations	xxi
Chapter 1 Introduction	1
1.1 The Real-World Data and Challenges of DCNNs in Medical Image Segmentation	3
1.2 Proposed Solution	5
1.3 Thesis Structure	7
1.4 Publications	8
Chapter 2 Background and Motivation	9
2.1 Medical Imaging and Segmentation	9
2.1.1 From Pixels to Prognosis: Medical Imaging	11
2.1.2 Medical Image Segmentation: Historical Perspective .	13
2.2 Deep Learning in Medical Image Analysis	15
2.3 Deep Learning for Medical Image Segmentation	17
2.3.1 Evolving Techniques	19
2.3.2 Medical Image Interpretation and Intricacies	28
2.4 Precision in Medical Data Annotation	29
2.4.1 Deep learning Confronting Inconsistencies	30
2.5 Crafting the Bespoke Loss Function	31

Chapter 3	Uncertainty Map Interpretations in Medical Image Segmentation	34
3.1	Context	34
3.2	Introduction	35
3.3	Background	36
3.3.1	Bayesian Neural Networks	38
3.3.2	The Epistemic and Aleatoric Uncertainty in Medical Image Segmentation	43
3.3.3	Factors Influencing Uncertainty Map Interpretations	50
3.4	Methodology	65
3.4.1	Encoder-decoder DCNN for Image Segmentation . . .	69
3.4.2	Uncertainty Attention Map	70
3.4.3	Error-Weighted Loss	72
3.4.4	Aleatoric Loss	74
3.4.5	Combination Loss Functions	75
3.5	Experiments	75
3.5.1	Datasets	75
3.5.2	Results	79
3.6	Discussion	91
3.7	Conclusion	93
Chapter 4	Uncertainty-Guided and Spatially-Constrained Weakly Supervised Deep Medical Image Seg- mentation	96
4.1	Context	96
4.2	Introduction	98
4.3	Methodology	106
4.3.1	DCNN-based image segmentation	106
4.3.2	Uncertainty guided loss and background constraint .	108
4.3.3	CRF based spatial constraint loss	110

4.3.4	Total loss function	111
4.4	Experiments	111
4.4.1	Datasets	111
4.4.2	Baselines	112
4.4.3	Result	113
4.5	Discussion	116
4.6	Conclusion	121
Chapter 5	Understanding Multiple Sclerosis: Deep Con- volutional Neural Networks for Multiple Scle- rosis Lesion Segmentation, and Associated Challenges	123
5.1	Context	123
5.2	Introduction	125
5.2.1	Multiple Sclerosis	125
5.2.2	Multiple Sclerosis Supervised Lesion Segmentation . .	131
5.3	Multiple Sclerosis Data and Challenges	142
5.3.1	preprocessing Methods	142
5.4	Experiments	147
5.4.1	Preprocessing Pipeline Comparative Experiments . .	151
5.4.2	MS Lesion Segmentation Tailored Loss Function Com- parative Experiments	152
5.4.3	Understanding the Impact of Acquisition Site and Scanner on MS Lesion Segmentation Model Perfor- mance	154
5.5	Discussion	158
5.6	Conclusion	159
Chapter 6	Conclusions	161
	Bibliography	167

Appendices	205
Appendix A	Uncertainty Map Interpretations in Medical
	Image Segmentation
	206
Appendix B	Understanding Multiple Sclerosis: Challenges
	and Deep Convolutional Neural Network MS
	Lesion Segmentation
	210
B.1	Model Training and Parameter Settings 210

List of Tables

3.1	Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For ISIC2018 Dataset . . .	80
3.2	Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For BraTS Dataset . . .	80
3.3	Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For BUS17 Dataset . . .	80
3.4	Segmentation performance comparison in terms of dice coefficients for different models on ISIC2018, BraTS whole tumour and BUS17 datasets on a test subjects for baseline Unet, Unet with AL, Unet with UCF loss, Unet with ELF loss, Unet with the UCF loss and AL as regularisation term and Unet with ELF loss and AL as regularisation term. . . .	82
4.1	Experimentally assessed parameter values for the proposed method.	113
4.2	Skin lesion segmentation performance comparison on test set for baseline Full-ResUnet, Weak-ResUnet, GrabCut, DeepCut and UBC our proposed method.	113
5.1	Overview of MS lesion segmentation methods, highlighting their key characteristics and associated limitations.	141
5.2	Dataset details with the number of samples, minimum and maximum image dimensions, all with voxel dimensions of 1x1x1 millimetre.	142

5.3	The Training, Validation, and Test sets chosen from the three local datasets (ACTIMUS, SPMRC and MEDSCH) for the experiments. As can be seen, the Single-Site mixture is so-called as it uses data from only a single site in training and validation.	150
5.4	Summary of Dice score ranges for MS lesion segmentation models based on training and testing scenarios reported in the literature [Kamraoui et al., 2022]. The table shows the range of Dice scores achieved when models were trained on one dataset and tested on another (cross-dataset) versus when trained and tested on the same dataset. No statistically significant differences were observed in either scenario.	150
5.5	MS lesion segmentation performance (on 16 Test Subjects) comparison of preprocessing methods NP and UKBP for the baseline models DeepMedic and MUnet using Multi-Site Mixture data.	152
5.6	MS lesion segmentation performance comparison on tailored loss functions and the baseline models DeepMedic and MUnet using UKBP pipeline with Multi-Site mixture data.	153
5.7	MS lesion segmentation performance comparison on learning from Single-Site Mixture and evaluating on Images acquired from different site/scanner; Preprocessing using UKBP pipeline for the baseline models DeepMedic and MUnet and tailored loss functions.	157
A.1	Non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in ISIC2018 dataset. Bold value indicates that the result of Wilcoxon test is significant (significance rate is 0.01).	207

A.2	Non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in BraTS dataset. Bold value indicates that the result of Wilcoxon test is significant (significance rate is 0.01). . . .	208
A.3	Non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in BUS17 dataset. Bold value indicates that the result of Wilcoxon test is significant (significance rate is 0.01). . . .	209
B.1	Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For MS Dataset	211

List of Figures

2.1	The architecture of a K-layer encoder-decoder CNN with skip connections.	20
3.1	An example of MS Lesions in T2 FLAIR vs T1. Yellow arrows show a typical pathology in MS patients (white matter lesion) and how the lesion appears differently in T1 vs T2 FLAIR. T1 is excellent in showing grey and white matter, while T2 FLAIR is suitable for suppressing fluid signals and depicting tissue alteration such as white matter lesions hence the hyperintensity segments.	52
3.2	(a) a patient with MS MRI images with T1w and T2Flair sequences acquired on a 3T Achieva MR scanner (Philips Medical, Best, The Netherlands) using a 32 channel head coil. (b) a patient with MS MRI images with T1-weighted and T2 FLAIR sequences acquired on a 3T MR750 scanner (GE Healthcare, Milwaukee, USA) using a 32 channel head coil	53
3.3	Examples of high-resolution images in Breast Ultrasound Lesions Dataset (Dataset B) including Fibroadenoma, Invasive Ductal Carcinoma, Cyst, Papiloma, Lymph Node and Lymphoma.	54

3.4	Brain tumour images from the BraTS 2018 dataset depicting the noise in images causing low SNR. High SNR images offer clearer differentiation of tumour boundaries, whereas low SNR images introduce noise, complicating accurate segmentation.	56
3.5	Examples of common artefacts in dermoscopic skin lesion images, including dark corners, hair, color charts, rulers, measurement markers, and ink markers, as well as the presence of air bubbles due to interface fluids. These artefacts introduce aleatoric uncertainty by obscuring or altering lesion characteristics, complicating accurate segmentation and analysis. Understanding and mitigating these artefacts is crucial for reducing uncertainty and enhancing the reliability of diagnostic interpretations.	57
3.6	(a) Distribution of uncertainty in TP and FP pixels (generated from calculating epistemic uncertainty map at inference time), highlighting the model’s confidence variability in lesion segmentation, showing generally high epistemic uncertainty for both TP and FP pixels. (b) Visualisation of the model’s output for a large lesion, including the original image, ground truth, prediction, and uncertainty map, demonstrating the model’s response to lesion sizes beyond its training scope.	60
3.7	Histogram plot of the breast cancer lesion dataset categorised by lesion types. The tail of this histogram depicts the OOD samples including Unknown (UNK), Ductal Carcinoma In Situ (DCIS), Lymph Node (LN), Papiloma (PAP), Invasive Lobular Carcinoma (ILC), Lymphoma (LP) lesions.	62

3.8	How aleatoric uncertainty can become epistemic uncertainty in when the model trained on low quality data encounters high quality data with fine details that the model has not seen before. The lesions are contoured with red lines.	67
3.9	The overall architecture of the proposed models	69
3.10	Variation of the Uncertainty Weight Function for Different Values of β : These plots illustrate how the uncertainty weight function changes with the entropy $H[y x, \mathcal{D}]$ for a range of β values, given fixed α value of 0.1. Each line represents a distinct β value, demonstrating the influence of β on the weight function response to entropy. The starting point of each line, marked by a circle, corresponds to the value of α , highlighting the initial influence of α on the uncertainty weight.	72
3.11	Comparative analysis of various models on the BraTS dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value. Letter T represents only a specific region of the brain and part of the whole tumour.	84

3.12	Comparative analysis of various models on the BraTS dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value. Letter <i>T</i> represents only a specific region of the brain with the whole tumour. Letter <i>A</i> represents only a specific region of the brain with the apparent abnormality that was not annotated as tumour.	87
3.13	Comparative analysis of various models on the ISIC2018 dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value. Letter <i>A</i> represents specific regions with artefacts.	88
3.14	Comparative analysis of various models on the BUS17 dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value.	90
4.1	The differences between full annotation, points, scribbles, and bounding box annotation on a skin lesion dataset sample. Data sample is selected from ISIC2018 Skin Lesion Dataset.	99

4.2	(a) Proposal-based weakly supervised medical image segmentation strategy, that uses weak annotations to generate pseudo-annotations known as <i>proposals</i> . These labels are generated iteratively to refine the parameters of deep CNNs to mimic the full supervision. (b) Weakly supervised medical image segmentation strategy using a tailored loss function that does not require iterative refinement of pseudo annotations.	103
4.3	Overview of the proposed bounding box based weakly supervised framework. We incorporated predictive uncertainty estimation during the training process to help the model focus on learning from image areas where it is more confident in its predictions, while also considering a background constraint. Furthermore, we included a local spatial constraint based on CRF in the loss function to enforce consistency in the predicted labels within a specific region.	107
4.4	Plot comparing the distribution of Prediction Dice Coefficient scores for different methods (Proposed Method, DeepCut, and GrabCut) across the test set. The x-axis represents ranges of Dice Coefficient scores achieved for test set, while the y-axis shows the number of test images falling within each score range, expressed as a fraction of the total number of test images. This visualisation illustrates the frequency with which each method achieves various levels of Dice Coefficient scores on the test images.	114
4.5	Selected DeepCut test results contouring the predicted lesion region over the original image on the models from epoch 1 to epoch 50 with step size of 10 (left to right), showing the vanishing issue of pseudo-masks during training.	115

4.6	Example skin lesion segmentation results for all compared methods: from left to right (a) original image, (b) ground truth, (c) Full-ResUnet, (d) Weak-ResUnet (e) GrabCut, (f) DeepCut (g) Proposed method.	117
4.7	The proposed UBC method progress during training from epoch 1 to epoch 20 on 2 sample images from the ISIC2018 dataset: (a) Original Image; (b) Ground Truth; (c) Bounding Box Annotation; (d) Prediction Result; and (e) Uncertainty Heatmap with red as the highest uncertainty value and dark blue as the lowest uncertainty value.	119
5.1	(a) Presents T1 and T2 FLAIR brain MRI scans of an MS patient, showcasing a slice where no lesions are detected. (b) shows T1 and T2 FLAIR brain MRI scans of another MS patient, highlighting a slice where a lesion is clearly visible in yellow in both the coronal and axial views.	129
5.2	Axial, sagittal, and coronal 3T FLAIR* images of relapsing–remitting multiple sclerosis; showing MS lesions with central vein sign [Maggi et al., 2018]	130
5.3	The original T1, T2 FLAIR and the resulting T1, T2 FLAIR and mask volumes by using the NP procedure. One subject per dataset are chosen for illustration, from top to bottom, subjects are selected from ACTIMUS, MEDSCH and SP-MMRC datasets, respectively. For each image, the Axial, Coronal and Sagittal views are illustrated.	144
5.4	The NP results on some of the challenging T2 FLAIR images of MEDSCH and SPM-MRC datasets (from top to bottom). Parts of the brain tissue are not captured by NP as marked by the yellow arrows.	144

5.5	The original T1, T2 FLAIR and the resulting T1, T2 FLAIR and mask volumes by using the UKBP procedure. One subject per dataset are chosen for illustration, from top to bottom, subjects are selected from ACTIMUS, MEDSCH and SPM-MRC datasets, respectively. For each image, the Axial, Coronal and Sagittal views are illustrated.	146
5.6	Ground truth alignment; the top row shows slices of three samples from their original volumes with and without contoured ground truths. The bottom row shows the respective results from The Modified UKBP preprocessing with and without contoured ground truths.	147
5.7	The comparison between NP and UKBP; each row shows the original, NP and UKBP T2 FLAIR slices of a subject, respectively.	147
5.8	Comparative visualisation of predictive performance across three distinct test samples: MUnet employing baseline cross-entropy loss (Baseline), MUnet with the proposed tailored loss functions, and DeepMedic, all predictions are overlayed on T2 FLAIR images, displayed in axial view. Each row corresponds to the results from a single sample, showcasing the efficacy of each approach.	156
5.9	Comparative visualisation of predictive performance across three distinct test samples: MUnet with the best proposed tailored loss functions (UCF and UCF+AL), and DeepMedic, all predictions are overlayed on T2 FLAIR images, displayed in coronal view. Each row corresponds to the results from a single sample, showcasing the efficacy of each approach. . . .	157

Abbreviations

- 2D** two-Dimensional.
- 3D** three-Dimensional.
- 4D** four-dimensional.
- ACU-Net** Attention Context U-Net.
- AI** Artificial Intelligence.
- AL** Aleatoric Loss.
- ANN** Artificial Neural Network.
- BET** Brain Extraction Tool.
- BMA** Bayesian Model Average.
- BNN** Bayesian Neural Networks.
- BraTS** Brain Tumour Segmentation challenge.
- BUS17** Breast Ultrasound Lesions dataset.
- CAMs** Class Activation Maps.
- CNN** Convolutional Neural Networks.
- CRF** Conditional Random Fields.
- CSF** CerebroSpinal Fluid.
- CT** Computed Tomography.
- CVS** Central Vein Sign.
- CYST** Unspecified Cysts.
- DAC** Dense Atrous Convolution.
- DCIS** Ductal Carcinoma In Situ.

DCNN Deep Convolutional Neural Networks.

DLB Deep Lesion Brain.

ELF Error-focused Loss Function.

EPI Echo-Planar Imaging.

FA Fibroadenomas.

FC Fully Connected.

FLIRT FMRIB's Linear Image Registration Tool.

fMRI Functional Magnetic Resonance Imaging.

FOV Field Of View.

FP False Positive.

FSL FMRIB Software Library.

GBM/HGG Glioblastoma.

GMM Gaussian Mixture Model.

GT Ground Truth.

IDC Invasive Ductal Carcinomas.

ILC Invasive Lobular Carcinoma.

ISIC2018 ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection challenge.

KL Kullback-Leibler.

LA Laplace Approximation.

LGG Lower Grade Glioma.

LN Lymph Node.

LP Lymphoma.

MALF Multi-Atlas Label Fusion.

MAP Maximum A Posteriori.

MC-Dropout Monte Carlo Dropout.

MCMC Markov Chain Monte Carlo.

MIL Multiple Instance Learning.

ML Machine Learning.

MLE Maximum Likelihood Estimation.

MP Magnetisation Preparation.

MPRAGE magnetisation-prepared rapid gradient-echo imaging.

MRF Markov Random Field.

MRI Magnetic Resonance Imaging.

MS Multiple Sclerosis.

MUnet Multi modality Unet.

NIFTI Neuroimaging Informatics Technology Initiative.

NLM Non-Local Means.

NP Naive Preprocessing.

OOD Out-Of-Distribution.

OS Overall Survival.

PAP Papiloma.

PET Positron-Emission Tomography.

RAGE RApid Gradient-Echo.

ReLU Rectified Linear Unit.

RF Random Forest.

RGB Red Green Blue.

RMNMS Rotation-invariant Multi-contrast Non-Local Means Segmentation.

RMP Residual Multi-Kernel Pooling.

ROI Regions Of Interest.

SNR Signal-to-Noise Ratio.

SPECT single photon emission computed tomography.

T1 T1-weighted magnetisation-prepared rapid gradient-echo imaging.

T2 FLAIR T2 Fluid-Attenuated Inversion Recovery.

TDA topological Data Analysis.

TOADS Topology-Preserving Anatomical Segmentation.

TP True Positive.

UCF Uncertainty-Calibrated loss Function.

UKBP UK Biobank Preprocessing Pipeline.

UNK Unknown.

VI Variational Inference.

WCE Weighted Cross Entropy.

WM White Matter.

X-Ray X-Radiation.

Chapter 1

Introduction

Medical imaging, a crucial component of radiological science, enables visualisation of the human body's anatomical structures and metabolic information, assisting in understanding anatomy, organ functions, and physiological processes. It is indispensable for both diagnosing and evaluating treatment responses [Dhawan, 2011].

Image segmentation plays a crucial role in medical image processing and has emerged as a key research area for understanding images. In image segmentation, an image is divided into several segments based on similar characteristics, essentially isolating the subject from its background. Commonly used imaging modalities include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron-Emission Tomography (PET), X-Radiation (X-Ray), and ultrasound, alongside RGB images such as microscopy, dermoscopy, and retinal images [Liu et al., 2021]. These medical images contain vital information for the diagnosis of patient conditions and are increasingly becoming a key component of clinical diagnosis.

Medical image segmentation involves using computer image processing to

analyse two-Dimensional (2D) or three-Dimensional (3D) images for segmentation, extraction, and 3D reconstruction of human organs and tissues, enhancing the precision and reliability of medical diagnoses [Liu et al., 2021]. This process involves several stages, including data collection, pre-processing, the application of the segmentation method, and performance evaluation, to improve lesion analysis and organ delineation for better diagnostic outcomes.

Deep learning is extensively employed in medical image analysis, spanning a range of anatomical areas such as the brain, eye, chest, digital pathology, microscopy, breast, cardiac, abdomen, and musculoskeletal regions. The applications of deep learning encompass tasks like image registration, classification, detection, and segmentation [Litjens et al., 2017]. Convolutional Neural Networks (CNN) stand out as one of the most effective and popular models in the realm of deep learning. Typically, CNNs are composed of three distinct types of layers: 1) convolutional layers, which utilise a set of weights, known as a kernel or filter, to convolve across the image and extract features; 2) nonlinear layers, where an activation function is applied to the feature maps, enabling the modeling of complex nonlinear relationships; and 3) pooling layers, which simplify the feature map by summarising small areas with statistical data like the average or maximum value [LeCun et al., 1998]. The connections between neurons in these layers are local, meaning each neuron is influenced by inputs from a nearby area, referred to as the receptive field, from the preceding layer. By organising these layers into a hierarchy, CNNs can learn to recognise features from progressively larger areas of an image. A key benefit of CNNs is their efficient use of parameters, achieved by sharing weights across all receptive fields within a layer [Minaee et al., 2021], which reduces the total number of parameters compared to fully-connected networks. In this way, Deep

Convolutional Neural Networks (DCNN) are well suited to working with image data, and so are increasingly popular in medical imaging.

1.1 The Real-World Data and Challenges of DCNNs in Medical Image Segmentation

Medical image segmentation using DCNNs holds immense potential, but faces multifaceted challenges. Achieving precision is important since segmentation targets pixel or voxel levels of accuracy; however, medical image boundaries can be difficult to identify. This intricacy is further complicated by rigorous preprocessing demands and frequent model parameter adjustments to ensure satisfactory accuracy. Complicating the landscape further, there is inconsistency in medical imaging equipment standards. The varying performance of CT and MRI machines can necessitate situation-specific deep-learning models [Yin et al., 2022]. Moreover, the sample size disparity between negative and positive sample regions can negatively affect segmentation accuracy. Despite the need for interpretable DCNNs, which can provide justification or confidence of their predictions in medical imaging, current models often lack clarity. Ensuring that medical professionals can comprehend, trust, and consequently utilise these technologies requires models to yield segmentation results that balance interpretability, reliability, and accuracy [Yin et al., 2022].

On the innovation front, fully automated segmentation via deep neural networks is of undeniable value. Still, it grapples with obstacles such as class imbalances, noise interferences, and the serious implications of missed detections [Wang et al., 2022]. Avenues for research include the design of network architectures accentuating encoder-decoder structures,

combined by skip connections, and task-centric feature extraction. Adopting bespoke loss functions becomes vital given medical imaging’s unique challenges. Transfer learning, built on pre-trained models, emerges as a potential remedy for noisy or sub-optimally labeled medical datasets. Additionally, interactive segmentation utilities anchored in deep learning can optimise both the efficacy and precision of user inputs [Wang et al., 2022].

Interpretability in clinical decision-making is paramount for physicians to trust Machine Learning (ML) predictions, particularly in disease diagnosis. Interpretable ML provides insights into domain-specific problems by extracting relevant data-based knowledge. Interpretation methods fall into two categories: model-based and posthoc interpretability. The model-based methods constrain the model to offer insightful information on learned relationships. The latter method delves into the relationships a model has learned, evidenced by studies evaluating feature importance scores. Another crucial aspect is Uncertainty Quantification, which gives a confidence measure to model predictions, acting as a form of posthoc interpretability. Recent works have focused on quantifying uncertainty in deep learning applications for medical imaging, such as lesion detection and disease classification. There is also growing interest in considering label uncertainty during network architecture design and training [Zhou et al., 2021].

Deep learning has significantly improved the accuracy of medical image segmentation due to its capability to handle complex scenarios. However, it requires a large number of labelled samples for training. Obtaining such extensive datasets for medical imaging is not easy, and labelling new images is both expensive and time-intensive. To overcome this, various methods have been explored. These include data augmentation, which involves expanding the training dataset by artificially modifying images; transfer learning, which uses models from related domains; and patch-wise training, which

splits images into smaller sections. Another important approach is weakly supervised learning, which is particularly beneficial when there is limited or poorly labeled data, utilising various forms of weak annotations such as image-level labels, sparse points, scribbles, and bounding boxes. These approaches have shown promise in training effective models, even without strong annotations.

1.2 Proposed Solution

This thesis aims to address the above-discussed two fundamental challenges in deep medical image segmentation: 1) the output reliability and interpretability problem, and 2) the strong annotation scarcity problem. We do this by exploring tailored loss functions for deep medical image segmentation. To this end, the following section presents a brief summary for the reader of the strategic utilisation of our proposed tailored loss functions, towards tackling these challenges for various medical datasets and the application of some of our proposed methods, on a specific challenging domain:

Uncertainty-Aware Loss Function for Fully Supervised Medical Image Segmentation: The aim was to improve the interpretability of uncertainty estimation in the context of medical image segmentation, a task complicated by the nature of the data and inherent uncertainties within it. This thesis explored the use of Bayesian Deep Convolutional Neural Networks, emphasising the development of specific loss functions designed to handle uncertainty more effectively. These included the epistemic uncertainty attention map on cross-entropy loss, an error feedback weight map on cross-entropy loss, and combinations of these with aleatoric loss as a

regularisation term.

Our research underscored the importance of integrating tailored uncertainty management strategies into medical diagnostics. It demonstrated that different loss functions manage uncertainty in diverse ways, influencing the accuracy and consistency of the interpretations derived from uncertainty maps. This careful approach to managing uncertainty is crucial for the advancement of medical diagnostics, aiming to enhance the reliability of automated systems and, consequently, improve patient care and treatment outcomes.

Uncertainty-Guided and Spatially-Constrained Weakly Supervised

Deep Medical Image Segmentation: The objective of this work was development and evaluation of a new method for segmenting skin lesions in images using only bounding box annotation for training data. The method employs a unique combination of techniques to improve the accuracy of segmentation. It introduces a customised loss function that incorporates both a weighted cross entropy, by uncertainty estimation, and a regularisation term based on Conditional Random Fields (CRF) to enhance segmentation based on the image’s details, beyond just the initial rough annotations.

MS Lesion Segmentation Case Study: The objective of this case study was to present the complexities of MS lesion segmentation, specifically focusing on the challenges posed by the intricate nature of MS. The case study is focused on how variations in image acquisition methods and the design of tailored loss functions can play a pivotal role in addressing these challenges, particularly when dealing with data that deviates from what the model has previously encountered. To investigate this, we explore the novel application of tailored loss function models on in-house acquired,

and highly challenging MS datasets.

Part of the case study explores the impact of preprocessing techniques on the segmentation of brain tissue from MRI images, emphasising that the choice of preprocessing pipeline can greatly affect the segmentation results.

This case study further compares the effectiveness of the loss functions in improving segmentation accuracy specifically for the unseen data acquired from different scanners and across different sites. The proposed loss functions showed comparable performance to methods that focused on customising the architecture of the DCNN, highlighting the importance of selecting the right loss function.

1.3 Thesis Structure

The thesis outline is as follows:

- Chapter 2 provides a background of medical imaging and specifically focusing on medical image segmentation, alongside the deep convolutional neural networks employed in the realm of medical image segmentation. This chapter further discusses the particular challenges of medical imaging, and precision of annotations in this field. Finally, the chapter discusses the crucial role of loss functions in the performance of convolutional neural networks in medical image segmentation.
- Chapter 3 presents an uncertainty-guided DCNN segmentation method. This chapter proposes uncertainty-aware loss functions and evaluates the interpretation of the uncertainty estimation measures at test-time when the model is trained with this uncertainty-aware loss.

- Chapter 4 introduces a novel loss function for weakly supervised medical image segmentation methods using spatial constraints and uncertainty-aware methods during training. The proposed model used bounding box annotations as ground truth.
- Chapter 5 first reviews standard techniques in segmentation of multiple sclerosis lesions in MRI images. This chapter also covers the challenges of MS data and examines the impact of preprocessing techniques. Finally, this work evaluates the performance of the proposed fully supervised segmentation solutions on MS lesion segmentation as a particularly challenging case study.
- Chapter 6 concludes this thesis and provides a summary of the main contributions in the realm of medical image segmentation, model prediction reliability and weakly supervised medical image segmentation.

1.4 Publications

Golnar Khalili Zadeh Mahani, Xin Chen, Nikolaos Evangelou, Stamatios Sotiropoulos, Paul S. Morgan, Andrew P. French *Uncertainty Map Interpretations in Medical Image Segmentation*, [Manuscript in preparation], (Extracted from Chapter 3).

Golnar Khalili Zadeh Mahani, Ruizhe Li, Nikolaos Evangelou, Stamatios Sotiropoulos, Paul S. Morgan, Andrew P. French, Xin Chen *Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially constrained loss*, 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) [[10.1109/ISBI52829.2022.9761558](#)], (Extracted from Chapter 4).

Chapter 2

Background and Motivation

2.1 Medical Imaging and Segmentation

Precise investigation and diagnosis are undeniably vital in the management of all medical conditions. Medical imaging, which involves creating visual representations of various tissues and organs within the human body, plays a crucial role in monitoring both normal and abnormal anatomical and physiological aspects of the body [Flower, 2012, Hussain et al., 2022].

Tracing the roots and evolution of medical imaging, and the complexities and nuances involved in distinguishing the truly groundbreaking innovations from the multitude of advancements in the field is challenging. Significant milestones in the development of medical imaging happened from 1895 to 1987. No major new imaging modalities have been introduced post-1988. This does not mean progress stopped, but rather that the foundational modalities had already been established by that time, with subsequent advancements being in the realm of technological details, uses, and applications [Flower, 2012].

From the sole use of planar X-Ray imaging in the early 20th century, the field has expanded to include advanced multidimensional imaging modalities such as CT, MRI, single photon emission computed tomography (SPECT), PET, and ultrasound. These modern techniques rely heavily on high-speed electronics and sophisticated computer technology for data acquisition, image reconstruction, and display. They allow for the visualisation of hidden diagnostic features which are hard or impossible to see with traditional planar methods, and these features can also be quantified and analysed to aid in diagnosis and treatment evaluation. While planar imaging methods, such as chest X-Rays and mammograms still hold clinical significance due to their cost-effectiveness and reliability, the multidimensional visualisation offered by newer methods delivers unparalleled clinical information, especially in critical applications [Dhawan, 2011].

Medical imaging encompasses a systematic five-step paradigm that begins with understanding the physiology of the object and its imaging medium. This understanding is crucial for selecting the appropriate imaging instruments and designing image processing techniques for accurate interpretation. This involves considering both static and dynamic properties of the biological tissue, such as tissue density or blood flow. The next step is understanding the physics of imaging, where the principle of data acquisition varies, such as the difference between X-Ray and MRI modalities. The third step involves the imaging instrumentation, where the quality of images relies on factors like signal-to-noise ratio, resolution, and artefact removal. Data acquisition and image reconstruction, the fourth step, emphasises optimising methods for the best spatial and temporal resolution, balancing between speed of acquisition and image detail, and employing suitable algorithms for image generation. Finally, the fifth step, image analysis and its applications, seeks to enhance diagnostic information

through various processing and analysis methods, supporting both manual and computer-assisted interpretations crucial for diagnosis and treatment monitoring [Dhawan, 2011].

2.1.1 From Pixels to Prognosis: Medical Imaging

In the realm of medical imaging, various techniques, such as transmission or nuclear resonance, can generate images of organs in both 2D and 3D formats, and even 4D format, which captures the dynamic changes in 3D organs over time. Despite the increasing complexity and sophistication of imaging methods, the fundamental principles governing how images are formed remain consistent. These foundational principles are universally applied, bridging the gap from basic 2D optical imaging to advanced multidimensional medical imaging systems.[Dhawan, 2011]. The essential commonality among these approaches is their dependence on fundamental principles of physics and optics, such as the interactions of light, sound, or magnetic fields with tissues. Nevertheless, the main distinctions emerge in the application and extension of these principles to attain the intended dimensional depiction. In 2D imaging, the emphasis is on capturing or displaying a 2D representation, whereas 3D and 4D imaging entail reconstruction algorithms and often necessitate incorporating temporal variations or supplementary dimensional information.

Imaging modalities are generally divided into two categories: anatomical (like X-Rays, ultrasound, and MRI) that differentiate bodily constituents such as water, bone, and soft tissue; and functional (like Functional Magnetic Resonance Imaging (fMRI), SPECT, and PET) that discern metabolism levels stemming from specific biochemical activities, possibly induced by radio-pharmaceutical substances [Dhawan, 2011]. fMRI, for

instance, is a technique that measures changes in blood flow and oxygenation levels in the brain, effectively unveiling neural activity. This technology has revolutionised our understanding of brain function and is particularly valuable in studying cognitive processes [Logothetis, 2008]. On the other hand, MRI T2 Fluid-Attenuated Inversion Recovery (T2 FLAIR) for example, is a specialised magnetic resonance imaging sequence that suppresses the signal from cerebrospinal fluid [Hajnal et al., 1992b], enhancing the visualisation of lesions and abnormalities within the brain [Hajnal et al., 1992a]. This technique is invaluable in diagnosing conditions such as multiple sclerosis and other neurological disorders by highlighting subtle structural changes in the brain.

Beyond generating various image modalities in medical imaging, the next aspect of medical imaging is analysing the images. The main goal of analysing these images is to enhance clinical research and treatment outcomes. Deep learning has transformed this analysis, excelling in tasks such as image registration, segmentation, feature identification, and categorisation. This transformation is primarily due to advanced computational capacities and the re-emergence of deep convolutional neural networks. Deep learning models excel at identifying underlying patterns in images, helping healthcare professionals make precise diagnoses. It has emerged as the main tool for tasks such as organ mapping, identifying cancers, classifying diseases, and aiding computer-driven diagnosis. Numerous deep learning strategies have been introduced for various diagnostic applications in medical imaging [Sistaninejhad et al., 2023].

2.1.2 Medical Image Segmentation: Historical Perspective

Medical image segmentation is a vital process in enhancing the visibility of anatomical or pathological structures, playing an indispensable role in numerous clinical applications, particularly in computer-aided diagnosis and smart medicine. By partitioning an image into distinct regions, it clusters neighbouring pixels based on predefined similarity criteria related to their features, categorising pixels to emphasise analogous areas [Despotović et al., 2015]. This technique is extensively used in segmenting various organs and abnormalities, such as the liver [Bilic et al., 2023], brain [Jyothi and Singh, 2023], cells [Greenwald et al., 2022], lungs [Teixeira et al., 2021], and heart [Chen et al., 2020a]. Its significance in radiological assessments and diagnoses has led to the development of a multitude of segmentation techniques, each with varying accuracy and complexity levels. Image segmentation plays a vital role in medical tasks, where it assists in treatment planning and ensures minimal damage to healthy tissue. Furthermore, it aids in volumetric analysis which is crucial for diagnostics and clinical decisions. Segmentation techniques have historically evolved from manual methods to today's more automated methods, aided by technological advancements [Trimpl et al., 2022]:

- **Manual Contouring:** Manual contouring involves clinicians manually outlining specific regions on medical images to define various anatomical structures or disease areas. This process entails segmenting the image dataset slice by slice, demanding a considerable level of proficiency and anatomical knowledge. The delineated contours serve multiple functions, including tumour volume calculation, radiation therapy planning, and disease progression evaluation. Manual Contouring

gives clinicians complete control but is tedious, time-consuming, and subject to variability among different observers

- **Interactive Segmentation:** This bridges the gap between manual and fully automatic methods. The approach requires user input but minimises it by leveraging predictions from prior user modifications. Interactive Segmentation methods offer more control and consistency than fully manual methods, requiring fewer user interactions. The user can either generate contours from scratch or refine the predicted segmentations. However, the surge in interest towards fully automatic methods has overshadowed interactive techniques, presenting an opportunity for advancements in this domain.
- **Fully Automatic Segmentation:** With the progression in machine learning, these approaches have become quicker and more consistent. Recent advancements of these methods have led to delivering outputs comparable in quality to manual procedures. Nonetheless, they come with obstacles: they mostly require extensive, manually-labelled data sets; encounter challenges owing to the inconsistency of certain anatomical features; and are affected by the variation in scanners and image acquisition protocols. Moreover, even with the precision such models provide, they still require expert review of the model output, due to the complexity of model architectures specifically DCNN models.

Image Segmentation involves dividing an image into distinct regions by clustering neighbouring pixels based on a predefined similarity criterion rooted in pixel properties or features. Essentially, it is a method of pixel categorisation to highlight similar areas within an image. From this perspective, segmentation techniques can be categorised into three main types:

edge-based methods that use edges to define object boundaries [Dhawan, 2011], pixel-based methods utilising image histogram statistics [Lee et al., 1990, Mittal et al., 2021], and region-based methods [Adams and Bischof, 1994, Chang and Li, 1994] where pixels are grouped directly based on similarity criteria.

The realm of medical image segmentation has seen significant progress thanks to the introduction of diverse methods and techniques. The evolution of deformable models [Xu et al., 2000], region growing [Adams and Bischof, 1994], clustering [Coleman and Andrews, 1979], classifiers [Rohlfing et al., 2004], Bayesian approaches [Chang et al., 2008], neural networks [Moghaddam and Soltanian-Zadeh, 2011, Sultana et al., 2020], thresholding [Mardia and Hainsworth, 1988], and other innovative strategies underscore the notable advancements in this area.

2.2 Deep Learning in Medical Image Analysis

DCNNs have revolutionised the field of medical imaging, offering numerous applications. Its potential for rapid and accurate image analysis can especially benefit diagnosis and treatment in various clinical areas. While traditional MRI analysis is time-consuming, requiring hours to analyse scans, DCNNs can streamline this process, optimising the arrangement of billions of voxels and accelerating image processing for radiologists. Additionally, DCNNs can enhance image segmentation, a crucial step that identifies tissue types. A trained DCNN can provide this segmentation quickly and efficiently, and tools like 2D and 3D DCNNs can predict segmentation maps either slice-by-slice or using volumetric data [Prabha et al., 2022].

DCNNs applications encompass tasks such as image registration, classification, localisation, and segmentation, as highlighted in the survey by [Litjens et al., 2017]. In this section, we briefly explain some of these applications.

Medical Image Registration In medical image analysis, registration [Wu et al., 2013, Yang et al., 2016, Simonovsky et al., 2016, Cheng et al., 2018], or spatial alignment, is a routine task that involves determining a coordinate transformation from one image to another. The most common approaches in the state-of-the-art are: (1) employing deep learning networks to estimate the similarity between two images, facilitating an iterative optimisation process, and (2) directly forecasting the transformation parameters through deep regression networks.

Medical Image Classification Deep learning first made significant strides in medical image analysis through image classification [Brosch et al., 2013, Menegola et al., 2016, Kawahara et al., 2017]. Here, one or multiple images serve as input to determine a singular diagnostic outcome. Lesion classification [Kawahara and Hamarneh, 2016, Nie et al., 2016, Xu et al., 2014] in medical imaging typically concentrates on categorising a specific, pre-identified section of the image into multiple categories. Achieving accurate classification often necessitates both local data about the lesion’s appearance and broader contextual information about its position.

Medical Image Localisation Another application of DCNNs are in organ, region and landmark localisation [Kumar et al., 2016, De Vos et al., 2016]. Identifying the location of anatomical objects [Lo et al., 1995, Teramoto et al., 2016], whether spatially or temporally, such as organs or landmarks, is crucial for pre-processing in segmentation or during the clin-

ical process for therapeutic planning and intervention. Identifying objects or lesions in images plays a pivotal role in diagnosis and often demands significant effort from clinicians. These tasks generally involve pinpointing and recognising tiny lesions within the entirety of an image.

Medical Image segmentation Segmenting organs and other structures in medical images facilitates quantitative evaluations of clinical attributes, such as volume and shape. This segmentation often serves as a foundational step in computer-aided detection processes [Prabha et al., 2022, Litjens et al., 2017]. As image segmentation is the main focus of this thesis, we discuss this topic further in the following sections.

2.3 Deep Learning for Medical Image Segmentation

To assist clinicians in precise diagnoses, it is vital to segment essential objects in medical images and derive features from these segmented areas. Essentially, segmentation involves identifying the voxels/pixels that form the contour or inside of the target object. Deep learning applications in medical imaging frequently address segmentation, leading to a diverse range of methods, including the creation of distinct CNN-based architectures [Litjens et al., 2017]. CNNs are a category of deep learning models specifically designed for handling data with a grid-like structure, such as images. CNNs excel at learning spatial feature hierarchies, transitioning from simple to complex patterns.

CNNs primarily consist of a series of layers that each perform specific operations on the input data to extract features and ultimately perform tasks

such as classification, detection, or segmentation. The main types of layers found in CNNs for segmentation are *Convolutional Layers*, *Activation Layers*, *Downsampling Layers*, *Upsampling Layers*, *Normalisation Layers*, and *Dropout Layers*.

The core building blocks of a CNN, *convolutional layers* apply a set of learnable filters (or kernels) to the input. These filters slide across the input image dimensions to produce feature maps, capturing local dependencies and patterns such as edges, textures, or more complex shapes in higher layers. The convolution operation helps preserve the spatial relationship between pixels. Following the convolutional layers, *activation functions* are applied to introduce non-linearities into the model, allowing it to learn more complex patterns. *Downsampling layers* reduce the spatial size of the feature maps, making the model more computationally efficient and robust to variations in the input. *Upsampling Layers* are used to increase the spatial dimensions of feature maps, allowing the network to construct a detailed, high-resolution segmentation map. *Normalisation Layers* are commonly used in CNNs to stabilise and accelerate training. *Dropout Layers* are used as a regularisation technique to prevent overfitting by dropping a set proportion of neurons in the network during training.

CNNs have become the focal point of research in image processing and computer vision due to their capability to hierarchically represent image features. Given their resilience to factors such as image noise, blur, and contrast, CNNs have consistently delivered better segmentation outcomes for medical imagery [Wang et al., 2022].

2.3.1 Evolving Techniques

In medical image segmentation, modern architectures largely adopt encoder-decoder networks. Consider the encoder-decoder network depicted in Figure 2.1. The encoder role is to transform an input signal (such as an image) into a set of features, which are essentially a more compact and abstract representation of the original input. This process involves several layers that progressively compress the input data and extract key features. The decoder then takes these compressed features and reconstructs the output, which often resembles the input but may be modified in terms of resolution or additional details.

The architecture is symmetric, meaning the encoder and decoder have the same number of layers. Each encoder layer reduces the data dimensionality through operations like convolution (to convolve across the image and extract features) and pooling (to summarise data). After these operations, the data is passed through an activation function, which enables the modeling of complex nonlinear relationships. Conversely, each decoder layer increases the data dimensionality by applying unpooling (which restores data size) followed by convolution, thereby reconstructing the data to its original or a modified form. This symmetry ensures that the transformations applied by the encoder are effectively reversed by the decoder [Ye and Sung, 2019].

Skip connections are a key feature in these networks. They allow the output from an encoder layer to bypass intermediate layers and connect directly to the corresponding layer in the decoder. This approach helps preserve important details that might otherwise be lost during the encoding process. By combining the skipped information with the decoder's output, the network maintains important details and produces a more accurate

reconstruction [Ronneberger et al., 2015]. For a detailed mathematical explanation of the encoder-decoder networks with skip connection, the reader is referred to [Ye and Sung, 2019].

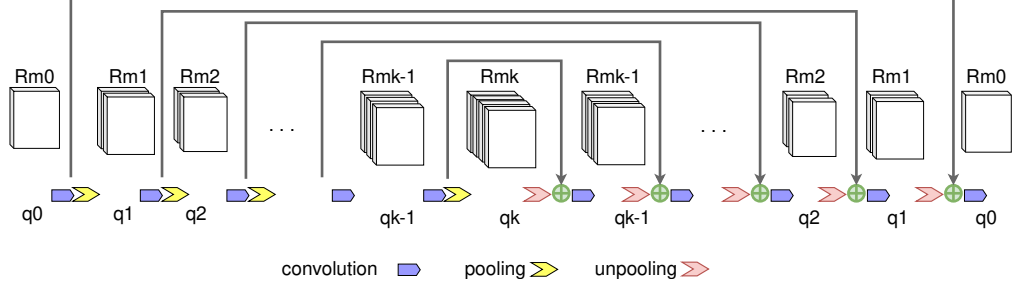


Figure 2.1: The architecture of a K-layer encoder-decoder CNN with skip connections.

One of the significant advancements in image segmentation from the medical imaging field was the development of the U-Net [Ronneberger et al., 2015] architecture that transformed the Fully Convolutional Networks into a U-type architecture combining both low- and high-level features through skip-connections [Salpea et al., 2022]. This architecture is essentially an encoder-decoder convolutional network with skip connections. Initially used for cell segmentation in light microscopy images, U-Net has demonstrated its capability to consistently and effectively learn relevant features across various medical image segmentation challenges [Zhou et al., 2021]. Subsequent innovations, such as the 3D U-Net [Çiçek et al., 2016] variant, VNet [Abdollahi et al., 2020], attention gate models [Oktay et al., 2018], and the CE-Net [Gu et al., 2019], Residual links [Ibtehaz and Rahman, 2020, He et al., 2016a], ResUnet [Zhang et al., 2018], further refined the capabilities of the base U-Net model.

The 3D U-Net architecture was proposed in [Çiçek et al., 2016], designed for volumetric image segmentation, to work with sparse annotations. The study provides two primary use cases: in a semi-automated scenario, users

annotate a few slices of a volume, and the network uses these annotations to generate a full 3D segmentation; in a fully-automated setup, the network is trained on a sparsely annotated dataset to segment new volumes without additional annotations. The 3D U-Net extends the traditional 2D U-Net by incorporating 3D operations such as convolutions and pooling, to handle volumetric data. The network’s performance was evaluated on the complex *Xenopus* kidney dataset, showing promising results for both semi-automated and fully-automated segmentation tasks. The study demonstrates that this approach is effective for learning from minimal annotated data and can be broadly applied to various biomedical imaging problems [Çiçek et al., 2016].

Attention U-Net proposed in [Oktay et al., 2018], is an enhancement of the standard U-Net architecture designed for medical image segmentation, particularly in handling structures with variable shapes and sizes. The core innovation is the integration of attention gates, which enable the model to focus on relevant regions of an image while suppressing irrelevant areas. This is achieved through a gating mechanism that applies attention coefficients to feature maps, selectively highlighting important features while minimising background noise. In the Attention U-Net architecture, attention gates are incorporated into the skip connections between the encoder and decoder paths. These gates refine feature maps by filtering out less relevant information before the concatenation step, thus improving the precision of segmentation results. The attention gates use a grid-based approach to generate attention coefficients specific to local regions, which contrasts with the global approach in previous models [Oktay et al., 2018]. This modification enhances the model’s ability to accurately segment small and variably-shaped structures without needing separate localisation modules. The Attention U-Net was tested on two large CT abdominal datasets,

demonstrating improved segmentation performance over the standard U-Net while maintaining computational efficiency [Oktay et al., 2018].

The ResUnet architecture was proposed for road extraction from aerial images [Zhang et al., 2018]. The architecture integrates the strengths of both residual learning and U-Net by replacing the standard convolutional units in U-Net with residual units to ease training and improve performance. The ResUnet architecture mirrors U-Net design with residual units that consist of a sequence of convolutional layers with batch normalisation and Rectified Linear Unit (ReLU) activations, along with identity mappings to facilitate better gradient flow and prevent degradation issues. The ResUnet includes three key components: encoding, bridging, and decoding. In the encoding path, residual units with strided convolutions reduce the feature map size while preserving crucial information. The bridging path connects the encoding and decoding stages, maintaining the features learned. In the decoding path, residual units are combined with up-sampled features from the encoding path through concatenation, enhancing the resolution and accuracy of the final segmentation map. ResUnet uses fewer parameters and eliminates the need for cropping, resulting in efficient processing with improved performance over U-Net and other methods [Zhang et al., 2018].

The CE-Net [Gu et al., 2019], introduces a novel approach for medical image segmentation that builds on the U-Net architecture to better preserve spatial information and capture high-level features. CE-Net comprises three main components: the feature encoder module, context extractor module, and feature decoder module. The feature encoder utilises a pretrained ResNet-34 [He et al., 2016a] block to improve feature extraction and network convergence. The context extractor includes a Dense Atrous Convolution (DAC) block and Residual Multi-Kernel Pooling (RMP) block [He et al., 2015]. The DAC block uses atrous convolutions [Chen et al., 2017]

with varying rates to capture features at different scales, enhancing the network ability to detect objects of various sizes. The RMP block employs multiple pooling kernels to gather contextual information from different scales and dimensions. The feature decoder module uses transposed convolutions and skip connections to restore high-resolution details and produce accurate segmentation masks. CE-Net demonstrated superior performance in various medical image segmentation tasks compared to traditional U-Net methods.

The VNet architecture was introduced as a deep convolutional neural network designed to improve road extraction from high-resolution remote sensing images [Abdollahi et al., 2020]. VNet modifies the traditional U-Net architecture by incorporating a dual-loss function, the cross-entropy-dice loss, to handle class imbalance and enhance segmentation accuracy. The VNet architecture is composed of two main parts: a compression path and a decompression path. The compression path reduces the spatial resolution through successive convolutional layers, where each phase includes residual connections that add the input of each phase to its final convolution output to ensure convergence. Unlike U-Net, VNet replaces max-pooling with convolutional layers to halve the feature map sizes, thereby maintaining higher resolution and better contextual information. The decompression path, on the other hand, progressively restores the original resolution of the feature maps using deconvolution operations followed by convolution layers, with the final layer applying a sigmoid function for probabilistic segmentation. This approach retains detailed information from the compression phase and improves segmentation performance [Abdollahi et al., 2020].

The MultiResUnet was proposed as an enhanced version of the classical U-Net architecture designed to improve performance in medical image segmentation, particularly in challenging scenarios [Ibtehaz and Rahman,

2020]. The U-Net architecture, known for its symmetrical encoder-decoder structure with skip connections, was effective but showed limitations when dealing with variations in object scales and discrepancies between encoder and decoder feature maps. To address these, MultiResUnet incorporated MultiRes blocks and Res paths [Ibtehaz and Rahman, 2020]. MultiRes blocks replace standard convolutional layers with a combination of 3×3 , 5×5 , and 7×7 convolutions, factorised to reduce memory usage while capturing multi-scale features more effectively. This block enhances the network ability to process features at various scales without excessive memory demands. Additionally, Res paths are introduced to bridge the semantic gap between encoder and decoder features by applying convolutional layers with residual connections along the skip paths, aligning features more closely and improving learning efficiency. The proposed MultiResUnet demonstrated better performance across multiple medical image datasets, particularly in complex or noisy images, achieving significant improvements over the original U-Net [Ibtehaz and Rahman, 2020].

Later versions of U-Net, such as U-Net++ [Zhou et al., 2018] and U-Net 3+ [Huang et al., 2020], and the nnU-Net [Isensee et al., 2018], and TransU-Net frameworks [Chen et al., 2021], have been developed to enhance accuracy and computational efficiency.

UNet++ [Zhou et al., 2018], was proposed as an architecture for medical image segmentation that enhances the U-Net model through redesigned skip pathways and deep supervision. Unlike U-Net, UNet++ incorporates nested, dense skip connections between the encoder and decoder networks. These skip pathways consist of dense convolution blocks that iteratively refine feature maps before they are merged with the decoder’s features, thereby minimising the semantic gap between encoder and decoder outputs. This approach ensures that the feature maps from both networks are more

semantically aligned, making the optimisation process more manageable [Zhou et al., 2018]. Additionally, UNet++ employs deep supervision, which involves supervising the network at multiple levels of the architecture. This provides finer-grained learning signals and improves segmentation accuracy. In practice, UNet++ achieved significant performance gains over both U-Net in various medical imaging tasks, including nodule, nuclei, liver, and polyp segmentation [Zhou et al., 2018].

The nnU-Net framework, designed as a self-adapting version of the classic U-Net architecture, simplifies and automates the process of medical image segmentation [Isensee et al., 2018]. The nnU-Net uses the core principles of U-Net, including its encoder-decoder structure with skip connections, but enhances its adaptability and performance across various datasets by focusing on automatic adjustments. The nnU-Net consists of three primary models: a 2D U-Net for 2D images, a 3D U-Net for volumetric data, and a U-Net Cascade that refines low-resolution segmentations from a 3D U-Net. This framework dynamically adapts to different image geometries and dataset sizes through automatic tuning of network configurations, pre-processing, and training parameters. The nnU-Net automatically handles preprocessing (such as normalisation, resampling), training (such as loss functions, data augmentation), inference (such as patch-based strategy and ensembling across test time augmentations and models) and a potential post-processing (the nnU-Net evaluation has shown automated, robust pre-processing and training strategies can significantly enhance segmentation performance without relying on elaborate network modifications) [[Isensee et al., 2018]].

The UNet 3+ architecture [Huang et al., 2020] extends the U-Net and UNet++ designs by utilising full-scale skip connections that link feature maps from different scales, allowing the network to combine fine-grained

details with high-level semantic information. Each decoder layer in UNet 3+ receives both smaller- and same-scale features from the encoder, and incorporates larger-scale features from within the decoder itself, improving the capture of position and boundary details. This is complemented by deep supervision, where side outputs from each decoder stage are supervised by ground truth, and a hybrid loss function. Additionally, UNet 3+ includes a classification-guided module that predicts the presence of organs in an image to reduce over-segmentation in non-organ areas. The approach demonstrates significant improvements in segmentation accuracy and computational efficiency on medical imaging datasets, surpassing previous methods in capturing detailed and accurate organ boundaries [Huang et al., 2020].

TransUNet [Chen et al., 2021] introduces a novel approach for medical image segmentation by integrating Transformers [Vaswani, 2017] with U-Net architecture. This hybrid model leverages the strengths of both architectures: Transformers provide robust global context through self-attention mechanisms, while U-Net contributes precise localisation with detailed spatial information. In TransUNet, the encoder uses a CNN to generate high-resolution feature maps, which are then tokenized into patches for processing by a Transformer. This combination allows the model to capture both global and local features. The decoder employs a cascaded upsampler that incrementally reconstructs high-resolution outputs from the encoded features, preserving spatial details through multiple upsampling steps and skip connections. This architecture enables TransUNet to overcome the limitations of pure Transformers, which can struggle with localisation due to their focus on global context. Empirical results show that TransUNet achieves promising performance in various medical image segmentation tasks, outperforming traditional CNN-based methods and other self-attention ap-

proaches [Chen et al., 2021].

Network function blocks are pivotal in enhancing the capabilities of CNNs for tasks such as medical image segmentation. Techniques like dense connections and inception structures [Szegedy et al., 2014] enhance performance by leveraging multi-scale convolution kernels without increasing the network’s depth. Depth-wise separable convolutions, as seen in MobileNet [Howard et al., 2017], are designed to minimise model size and memory footprint without major performance trade-offs. Additionally, attention mechanisms [Sun et al., 2020, Schlemper et al., 2019, Jetley et al., 2018] which assign varied weights to input features based on their relevance, have gained prominence for their efficacy in image segmentation tasks. All of these refinements have contributed to improving feature representation and boosting segmentation accuracy in CNNs [Wang et al., 2022].

The medical image segmentation field, despite benefiting from deep learning, grapples with several challenges. Dataset-related challenges encompass limited annotated datasets, class imbalances, sparse annotations, intensity inhomogeneities, and complexities in image texture. Addressing these, solutions such as image augmentation [Shorten and Khoshgoftaar, 2019], transfer learning [Weiss et al., 2016], and data synthesis are proposed. Regarding deep learning model challenges, remedies for these include dropout [Srivastava et al., 2014] techniques, model simplification, batch normalisation [Ioffe and Szegedy, 2015], careful initialisation [Arpit et al., 2019], and optimising parameters for computational efficiency [Malhotra et al., 2022].

2.3.2 Medical Image Interpretation and Intricacies

The interpretation of medical images, particularly MRI, is a complex task given the intricacy of delineating organs and lesions on a layer-by-layer basis. While manual methods offer a degree of precision, they are labour-intensive and prone to individual interpretation biases. In an attempt to streamline the process and mitigate subjectivity, automatic segmentation techniques have been introduced. However, these are not without their challenges, as medical imaging presents diverse modalities and each patient’s anatomical variations can differ significantly, complicating the task of achieving consistent and accurate segmentation. Thus, while technological advancements promise efficiency, there remains a need for further refinement to fully address the complexities inherent in medical image segmentation.

Medical imaging complexity presents unique challenges for deep learning solutions. The field possesses diverse imaging modalities with increasing pixel density, such as CT and MRI, which have achieved sub-millimeter spatial resolution. However, these images often vary significantly due to non-standardised acquisition protocols across different equipment, resulting in the "distribution drift" phenomenon. Although vast amounts of medical imaging data exist, they are scattered across various hospitals and imaging centres, making centralised open-source databases rare. Additionally, the variety of diseases visualised in these images is vast, but they exhibit a long-tailed distribution, meaning only a few diseases are common, while many are rare. Labeling these images is both resource-intensive and prone to inconsistencies, leading to noisy and sparse annotations. Furthermore, in the labelled sets, the data are often imbalanced, with, for example, tumour pixels being vastly outnumbered by normal tissue. Combining

these intricacies with the array of medical imaging tasks, such as detection, segmentation, and registration, results in a series of complex challenges to address [Zhou et al., 2021].

2.4 Precision in Medical Data Annotation

Creating precise manual segmentation masks, known as strong annotations, is time-consuming, especially for 3D images. To reduce the challenges associated with annotation, there has been a recent surge in research into weakly-supervised semantic segmentation [Tajbakhsh et al., 2020], which focuses on training segmentation models using various forms of weak annotations, including image-level annotations [Wu et al., 2019], sparse annotations including points [Kervadec et al., 2019, Qu et al., 2019], scribbles [Can et al., 2018], and bounding boxes [Rajchl et al., 2016, Yang et al., 2018, Kervadec et al., 2020]. Recent studies indicate that models can be effectively trained using weak annotations, even in the absence of strong ones [Lu et al., 2021].

Dealing with imprecise labels involves managing uncertainties or inaccuracies in segmentation mask edges, often found in medical imagery. These inaccuracies can stem from unintentional mistakes by annotators or differences in opinions among specialists, especially when interpreting unclear lesions. This noise can be simulated by modifying the vertices of polygons representing the labels, leading to segmentation errors. Another source of label noise is when the model uses its own predictions on unlabelled data, known as pseudo annotations. If not addressed, such noise can hinder the model’s performance. Strategies have been developed to address this, such as a weighted cross entropy loss function [Mirikharaji et al., 2019], which

controls each pixel’s contribution based on perceived annotation quality or iterative mask refinement [Min et al., 2019], often used in semi-supervised learning, wherein the model identifies noisy labels and reduces their impact during training [Tajbakhsh et al., 2020]. While techniques for handling noisy labels may either refine the noisy masks or leave them untouched, it is essential that they adaptively determine which pixels contribute to the loss, ensuring effective learning despite the noise [Tajbakhsh et al., 2020].

Learning with sparse labels refers to methodologies that deal with incomplete annotations, where only a portion of a 3D volume’s slices or a 2D image’s pixels are labelled such as points, scribbles, and bounding boxes. There are two predominant strategies [Tajbakhsh et al., 2020]: One is loss with mask completion to artificially reconstruct incomplete ground truth masks for training. For instance using active learning, and ranking non-annotated slices via the Dice similarity coefficient [Zhang et al., 2019b], or Grabcut [Rother et al., 2004] inspired methods to generate initial psuedo masks [Rajchl et al., 2016, Cai et al., 2018]. The other strategy is to improve the segmentation performance given sparse annotations without trying to fill in the missing mask pixels. For instance only the labelled pixels contribute to a weighted segmentation loss [Bokhorst et al., 2018] or using a class-balanced cross entropy loss [Çiçek et al., 2016] where only the labelled pixels contributed to loss.

2.4.1 Deep learning Confronting Inconsistencies

To effectively address the challenges of sparse and noisy labels in Deep Learning, there is a growing emphasis on approaches that efficiently utilise annotations. A number of learning methods have been utilised in the literature to address this problem, including transfer learning, domain adapta-

tion, self-supervised learning, semi-supervised learning, and weakly supervised learning. Refer to [Zhou et al., 2021] for a comprehensive review of the recent developments of these methods.

Inspired by the weakly supervised methods, I have dedicated Chapter 4 of this thesis to delve deeper into the nuances of effectively segmenting medical images while predominantly depending on bounding box annotations.

2.5 Crafting the Bespoke Loss Function

Loss functions serve a crucial role, especially in CNN-based segmentation methodologies by quantifying the disparity between the predicted segmentation and the actual ground truth. When constructing a CNN for segmentation purposes, one of the key steps involves designing an appropriate loss function. Throughout the training process, this loss function serves as a guiding principle, facilitating the network’s acquisition of meaningful predictions that closely align with the ground truth, as assessed by segmentation metrics like the Dice similarity coefficient. Additionally, the choice of loss function also governs how the network manages errors, such as false positives and false negatives. In the realm of medical image segmentation tasks, cross entropy and Dice loss have emerged as the most prevalent loss functions [Ma et al., 2021].

In medical image segmentation, various loss functions have been utilised. The cross-entropy loss, which computes the difference between the predicted and actual segmentation on a pixel-wise basis, is among the most prevalent. Due to challenges with class imbalance, weighted cross-entropy loss was introduced to adjust the value of positive and negative samples. The Dice loss is another metric, focusing on the overlap between segmen-

tation and ground truth. The Tversky Loss, a modified version of the Dice loss, and the Generalised Dice Loss was also introduced to address the class imbalance issues. Further enhancements include the boundary loss, which minimises the distance between segmented and labelled boundaries, and the exponential logarithmic loss, which combines focal loss and dice loss to cater to class imbalances. The design of these loss functions, although pivotal for improving segmentation results, remains task-specific, necessitating careful analysis for optimal outcomes [Wang et al., 2022].

In their work, [Ma et al., 2021] systematically and comprehensively categorise loss functions for CNN-based segmentation into five distinct groups based on their derivation to Distribution-based Loss, Region-based Loss, Boundary-based Loss, Compound Loss functions, and tailored loss functions ¹:

Distribution-based loss functions aim to minimise dissimilarity between two distributions, Cross Entropy and Focal loss being the most well known ones of this category.

Region-based loss functions aim to minimise the mismatch or maximise the overlap regions between ground truth and predicted segmentation. The popular Dice loss is the best known representative of this class of loss functions, followed by Jaccard loss, Tversky and focal Tversky loss.

Boundary-based loss, a relatively new type of loss function, aims to minimise the distance between ground truth and predicted segmentation and they normally are coupled with region-based loss function when being used for training neural networks. Hausdorff Distance loss is the most well-known one in this category.

¹Refer to [Ma et al., 2021] for a comprehensive review and analysis of Distribution-based Loss, Region-based Loss, Boundary-based Loss, Compound Loss functions.

Compound loss functions are the (weighted) combination between the loss functions, such as the combination of cross entropy and Dice loss or the combination of Dice loss with focal loss.

Tailored loss functions, in contrast to the versatile, ready-to-use loss functions suitable for a wide range of segmentation tasks, comprises loss functions specifically crafted for unique segmentation challenges.

Chapters 3, 4, and 5 of this thesis revolve around this particular category, with the majority of the thesis contributions dedicated to the development of tailored loss functions. These efforts are aimed at addressing the challenges introduced in the introduction by crafting well-designed loss functions.

Chapter 3

Uncertainty Map

Interpretations in Medical Image Segmentation

3.1 Context

Integrating uncertainty maps into medical image segmentation might provide valuable insights into the confidence of a model. However, the accuracy of these maps in conveying true uncertainty remains debatable. This chapter explores the intricacies of interpreting uncertainty maps in medical image segmentation, considering complicating factors such as lesion shape, border irregularity, inconsistencies in acquisition methods, variance in lesion texture, and the presence of image artefacts. We design experiments to study the interpretation of these maps and evaluate their reliability across different datasets. The contributions in this chapter are as follows:

- An in-depth analysis of how image characteristics influence uncer-

tainty map interpretation

- Assessment of the effect of the three proposed loss functions in uncertainty interpretation at inference time

3.2 Introduction

Medical image segmentation plays a pivotal role in clinical diagnosis and treatment. Deep learning models, often trained on datasets manually annotated by human experts, face challenges due to inherent variabilities and subjective assessments. This introduces ambiguities in predictions. To address this, techniques that combine deep learning methods with expert insight are essential, fostering more reliable human-computer interactions and enhancing decision support in clinical processes [Ching et al., 2018, Jensen et al., 2019].

Uncertainty estimation in healthcare involves measuring and understanding the inherent uncertainty or variability associated with medical predictions, diagnoses, and treatment outcomes [Seoni et al., 2023]. Uncertainty estimation is crucial to evaluate the reliability and effectiveness of deep learning models, especially in the realm of medical image analysis. With models often trained on large, yet noisy datasets, understanding the inherent variabilities becomes paramount. Especially in medical imaging, where decisions can have profound implications, understanding and quantifying uncertainties is not just beneficial but essential.

While uncertainty estimation is utilised as a means to protect decision support systems from vulnerabilities in CNN models, especially when tackling rare or challenging subjects [Ching et al., 2018, Bradshaw et al., 2017], its reliability remains a concern. Before utilising uncertainty maps to protect

against model vulnerabilities, their accuracy and interpretability must be subjected to examination. This chapter delves into this very aspect, exploring how factors such as lesion shape, border irregularity, the presence of image artefacts, variance in lesion texture, and acquisition methods might influence these maps.

In this chapter, we delve into the intricacies of interpreting uncertainty maps in medical image segmentation. We introduce novel uncertainty-aware loss functions for training a fully supervised DCNN model, with a focus on the role of uncertainty estimation during training. A key aim of this work is to investigate the potential influences of various image characteristics on uncertainty map interpretation, and the impact of three proposed tailored loss functions in refining uncertainty interpretation at inference time.

3.3 Background

Over the past decade, DCNNs have experienced significant advancements, leading to their integration into numerous research fields specifically including medical imaging. While their potential in high-risk medical fields is recognised, their deployment in practice remains limited. This limitation is attributed to factors such as the lack of transparency of a DCNN inference model, which makes it difficult to trust their outcomes, the inability to distinguish between in-domain and out-of-domain samples and the sensitivity to domain shifts, the inability to provide reliable uncertainty estimates and frequently occurring overconfident predictions. These issues often arise from inherent uncertainties in data or gaps in the knowledge of neural network model. To harness the potential of DCNNs in medical

applications, addressing these limitations necessitates the provision of accurate uncertainty estimates, which can aid decision-making in high-risk areas, especially where data sources are diverse or labelled data are sparse.

Recent research has shown growing interest in estimating uncertainty in DCNNs, typically by distinguishing between uncertainties caused by the model and those arising from the data. From data acquisition to developing a DCNN model and estimating uncertainty several potential factors and errors can influence DCNN predictions, including the variability in real world situations, the errors inherent to the measurement systems, the architecture specification of the DCNN and the training setting and parameters (such as batch size, optimiser, learning rate, regularisation, and weight initialisation) [Gawlikowski et al., 2023].

Techniques to estimate uncertainty in neural networks can be categorised into Single Deterministic, Bayesian and Ensemble methods. Single Deterministic methods utilise a single deterministic network and generate a prediction from a singular forward pass or by introducing augmented input data and generate several predictions from a singular pass [Gawlikowski et al., 2023]. Uncertainty in this category is either measured through supplementary techniques or directly predicted by the network itself. Bayesian methods encompass all stochastic DCNNs, meaning that consecutive forward passes with the same input might yield varying outcomes. In Ensemble methods, predictions are derived by combining results from multiple deterministic networks during inference [Gawlikowski et al., 2023]. For a comprehensive survey on Single Deterministic and Ensemble methods, readers are directed to references [Gawlikowski et al., 2023] and [Abdar et al., 2021].

In this chapter, we focus on the interpretability of uncertainty estimation

in Bayesian deep learning, specifically within the context of medical image segmentation. Consequently, this section offers a detailed review of the literature in the following topics: 1) main categories of Bayesian Neural Networks (BNN), 2) types of uncertainty in Bayesian modeling specific to medical image analysis, and 3) evaluating the trustworthiness of uncertainty estimation techniques in Bayesian deep learning applied to medical image analysis.

3.3.1 Bayesian Neural Networks

BNNs are a fusion of neural networks and Bayesian learning principles, representing a shift from traditional Artificial Neural Network (ANN) methods. While ANNs, such as feedforward networks, are constructed with input, hidden, and output layers to represent an arbitrary function $y = U(x)$, they traditionally focus on finding a point estimate of the network parameters using the backpropagation algorithm. This approach, often termed Maximum Likelihood Estimation (MLE) or Maximum A Posteriori (MAP) estimation when regularisation is applied, is straightforward with modern tools, but tends to lack transparency. ANNs may not be able to accurately predict outcomes when presented with data that are outside of their training distribution. Importantly, they are not able to indicate when they are unsure of an unfamiliar input. [Jospin et al., 2022].

BNNs, on the other hand, offer a paradigm shift. Instead of seeking a singular value for each parameter, BNNs infer a probability distribution over these parameters. This is achieved by computing the posterior distribution over the parameters based on both training data and a prior distribution. In this setup, each layer is not just a linear transformation followed by a nonlinear activation function but has stochastic components introduced,

either in the form of stochastic activations or stochastic weights. This inclusion of randomness effectively simulates multiple possible models with their associated probability distributions, making BNNs a form of ensemble learning.

How Do Bayesian Neural Networks Formulate Predictive Distributions?

As we mentioned above in traditional learning, the objective is to identify the optimal parameters of a model, usually by maximum-likelihood optimisation. Contrarily, in the Bayesian learning, after observing data $\mathcal{D}$, the learner discerns a posterior distribution $p(w|\mathcal{D})$ over the model parameters w . This distribution adheres to the Bayes rule:

$$p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w)$$

where $p(\mathcal{D}|w)$ is the likelihood of $\mathcal{D}$ for a model with parameters w , and $p(w)$ stands for the prior distribution over these parameters. The predictive distribution for a new test input (x) is then calculated by the Bayesian Model Average (BMA):

$$p(y|x, \mathcal{D}) = \int p(y|x, w)p(w|\mathcal{D}) dw,$$

with $p(y|x, w)$ being the predictive distribution given a specific parameter set w . In the Bayesian framework, the posterior over parameters of a modern neural network can encapsulate multiple coherent solutions to a given problem, representing different parameter configurations.

However, the exact computation of the BMA integral in the above equation is infeasible for neural networks, necessitating approximate inference techniques. The complexity of the approximation is due to the complex and high-dimensional posterior $p(w|\mathcal{D})$ [Izmailov et al., 2021].

Generally, classical training can be seen as a form of approximate Bayesian inference, which uses the approximate posterior $p(w|\mathcal{D}) = \delta(w = \hat{w})$ to calculate BMA. The Dirac delta function δ is zero everywhere except at $\hat{w} = \arg \max_w p(w|\mathcal{D})$. This results in the standard predictive distribution $p(y, x|\hat{w})$ [Wilson and Izmailov, 2020]. From this point of view, many suboptimal solutions will be preferable, including Gaussian posterior approximations for $p(w|\mathcal{D})$ even if the posterior or likelihood are actually highly non-Gaussian and have multiple peaks. The distinction between a classical and Bayesian approach is contingent upon the sharpness of the posterior $p(w|\mathcal{D})$. If the posterior is sharply peaked and the conditional predictive distribution $p(y|x, w)$ does not vary much in the areas where the posterior has mass, then there may be little to no difference, as a delta function may be a suitable approximation of the posterior for BMA [Wilson and Izmailov, 2020]. However, modern neural networks are usually highly underspecified by available data and therefore have diffuse likelihoods $p(\mathcal{D}|w)$, not strongly favouring any one setting of parameters. The likelihoods being diffuse and the parameters having a variety of possible settings, it is the ideal situation to use a Bayesian model average. This will result in an ensemble of many different models that are all highly effective, providing better calibration and accuracy than traditional training [Wilson and Izmailov, 2020].

In essence, a BNN can be understood as any stochastic ANN trained using Bayesian inference. The key advantage of BNNs is their ability to compute principled predictive uncertainty, which is invaluable in critical

applications. However, this comes with computational challenges, as estimating predictive uncertainty often requires tackling intractable integrals, necessitating approximation techniques like Monte Carlo Approximation. Beyond just uncertainty estimation, BNNs also open the door for powerful Bayesian tools in deep learning, enabling advanced features like model selection, compression, and active learning. BNNs can be divided into three categories [Gawlikowski et al., 2023]:

Sampling Methods

Sampling methods, often known as Monte Carlo methods, represent uncertainty without relying on a parametric model. Instead, they draw samples from a distribution to obtain non-parametric representations of the probability distributions. Markov Chain Monte Carlo (MCMC) is a popular method used with neural networks. The primary idea behind MCMC is to sample from arbitrary distributions by transitioning in a state space, where the transition is determined by a proposal distribution. Modern algorithms aim to make these methods scalable by using mini-batch versions of these methods, like Stochastic Gradient MCMC [Gawlikowski et al., 2023].

Laplacian Approximation

Laplace Approximation (LA) is a method that estimates the posterior distribution of neural network parameters by approximating it with a multivariate normal distribution centered around a local minimum of the loss surface (which corresponds to a peak of the posterior distribution). Specifically, LA builds a Gaussian approximation for the posterior around the MAP estimation of network weights. This means the mode of the Gaussian approximation is set to the mode of the posterior distribution and the

curvature of the log probability density is matched at that location [Murphy, 2012]. This approach relies on a second-order Taylor series expansion of the log posterior around the MAP estimate, which involves evaluating the first and second derivatives of the log-posterior.

One of the primary advantages of LA is that it offers post-hoc uncertainty descriptions, allowing for insights into the model’s performance without the need for re-training networks. The linearised LA method further enhances this by enabling sampling-free Bayesian inference. However, a central challenge with LA, especially when applied to modern deep neural networks, is the estimation of the Hessian matrix due to its large size. To overcome this, various strategies and approximations for the Hessian have been proposed. Despite these challenges, LA remains a versatile tool as it can be applied post-training and is compatible with networks using standard loss functions and activations. Recent advancements have further extended the applicability of LA, allowing it to accommodate large data sets, complex architectures, and real-time processing requirements [Gui et al., 2022, Gawlikowski et al., 2023, Abdar et al., 2021].

Variational Inference

Variational Inference (VI) is a technique that approximates the posterior distribution by optimising over a family of tractable distributions. The main objective is to make the approximation close to the desired posterior distribution. This closeness is measured using the Kullback-Leibler (KL) divergence. Variational methods have been extended and adapted to work efficiently with modern neural networks. VI has also been linked to existing stochastic elements in deep learning, like dropout, resulting in methods like Monte Carlo Dropout (MC-Dropout) [Gawlikowski et al., 2023, Abdar

et al., 2021, Gal and Ghahramani, 2016].

Among the methods used for uncertainty estimation, MC-Dropout stands out. It offers a computationally efficient way to estimate prediction uncertainty by incorporating dropout, a technique that prevents overfitting in neural networks. Despite the progress made, the main difficulty still lies in determining whether the uncertainties are genuine reflections of the model’s trustworthiness or simply the result of the technique used. In the field of medical imaging, many studies have used MC-Dropout for tasks such as segmentation [Abdar et al., 2021]. In this chapter, we focus on studies that utilised dropout as a Bayesian approximation and explore how various factors impact the interpretation of the resulting uncertainty estimates.

3.3.2 The Epistemic and Aleatoric Uncertainty in Medical Image Segmentation

Neural networks used for imaging are subject to a variety of factors, resulting in inherent uncertainty. These uncertainties are broadly classified into two categories: data uncertainty, referred to as aleatoric uncertainty and model uncertainty, commonly known as systemic or epistemic uncertainty. In the subsequent sections, we will delve deeper into the complexities of these two types of uncertainties, exploring their origins, implications, and the common techniques developed to estimate them in medical image segmentation. This deeper understanding is crucial, especially when considering the critical nature of medical image analysis and the need for reliable interpretations.

From the perspective of knowledge representation and decision-making,

differentiating between these two types of uncertainty is important, particularly in situations where the final decision can be postponed. For instance, a medical doctor who understands that their uncertainty regarding a patient’s illness stems from insufficient knowledge about the disease might choose to review the medical literature or seek advice from a colleague prior to making a decision [Senge et al., 2014].

The importance of accurately estimating uncertainty in deep neural networks cannot be overstated. Doing so is essential to address issues of confidence, interpretability, and robustness [Papernot and McDaniel, 2018]. Estimating uncertainty can help differentiate between aleatoric and epistemic uncertainties, increase trust in model predictions, and make the model more resilient to adversarial examples by ensuring the outputs are both reliable and interpretable.

Epistemic Uncertainty

Model uncertainty encapsulates the uncertainties attributed to limitations or inadequacies in the model. This could be due to flaws in the training process, an unsuitable model structure, or simply an insufficient number of training data [Gawlikowski et al., 2023].

One of the most popular stochastic regularisation techniques for deep learning models is dropout [Gal, 2016]. Dropout is used to avoid over-fitting of the network which can be interpreted as a form of approximate inference in a probabilistic interpretation of the model [Gal and Ghahramani, 2016, Srivastava et al., 2014]. When using the dropout as Bayesian approximation, in their study, Gal et al. proposed that a neural network, irrespective of its depth and nonlinearities, with dropout implemented before each weight layer, is analogous to an approximation of the probabilistic deep Gaussian

process [Gal and Ghahramani, 2016]. This finding holds true without any oversimplifying assumptions about dropout layers use, ensuring its applicability to any network architecture that employs dropout in real-world scenarios. The core objective of dropout, in this context, is to reduce the KL divergence [Kullback and Leibler, 1951] between a certain distribution and the posterior of a deep Gaussian process. This allows us to model the epistemic uncertainty using dropout. Predictive entropy can express a measure of ignorance about the actual structure of the model capturing the amount of information in the predictive distribution by:

$$H[y|x, \mathcal{D}] = - \sum_{c=0}^{c-1} p(y = c|x, \mathcal{D}) \log(p(y = c|x, \mathcal{D})) \quad (3.1)$$

where, H is the entropy of y given a sample image x and observed data $\mathcal{D}$ (referring to the training set) [Gal and Ghahramani, 2016, Pal and Pal, 1989]. The expected value of $p(y = c|x, \mathcal{D})$ is calculated by:

$$p(y = c|x, \mathcal{D}) = \int p(y = c|x, w) p(w|\mathcal{D}) dw \quad (3.2)$$

where w is the model parameters and the expectation can then be approximated with Monte Carlo estimation using the probability of x in class c with model parameters as the following:

$$p(y = c|x, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^M p(y = c|x, \widehat{w}_m) \quad (3.3)$$

for which, in practice, the probability vector p is calculated by passing the same input image x through the model M times [Gal, 2016], randomly dropping regions of feature maps. Therefore, the predictive entropy $H[y|x, \mathcal{D}]$ presents the epistemic uncertainty where the values range between 0 and 1. The higher the value of $H[y|x, \mathcal{D}]$, the more uncertain the prediction will be. The obtained epistemic uncertainty maps are normally used at

inference time in the literature.

Studies have been conducted to help experts better interpret predictions using epistemic uncertainty [Jensen et al., 2019, Teye et al., 2018, Gal and Ghahramani, 2016, DeVries and Taylor, 2018, Hu et al., 2019, Jungo et al., 2018a, Kampffmeyer et al., 2016]. Other research has used epistemic uncertainty to identify difficult [Jungo et al., 2018b] and adversarial samples [Feinman et al., 2017], suggesting further examination. Additionally, epistemic uncertainty estimation has been used to guide semi-supervised learning models [Yu et al., 2019, Sedai et al., 2019], especially when there is a large number of unlabelled samples and a limited amount of labelled data. These applications demonstrate the versatility and essential role of epistemic uncertainty in improving the accuracy and effectiveness of deep learning models.

Aleatoric Uncertainty

Data uncertainty arises directly from the data itself. It is the result of inevitable information losses when translating real-world phenomena into data samples. Aleatoric uncertainty is the variability in the outcome of an experiment that is due to inherent random effects in the data. An example of this type of uncertainty could be found in the coin flipping experiment; the results generated by each experiment have a stochastic element, which cannot be reduced by any additional source of information [Hüllermeier and Waegeman, 2021]. Hence, considering the best model that represents the coin flipping experiment could only give the possible probabilities of the outcome of this experiment being head or tail.

In classification, data uncertainty might arise from samples that do not provide sufficient clarity to confidently assign to a particular class. Such

information loss can arise from the limitations of measurement tools, such as image resolution limitations or from errors during data labeling [Gawlikowski et al., 2023].

In image segmentation, aleatoric uncertainty exhibits spatial correlation and can be heteroscedastic, since certain regions of an image might exhibit more uncertainty than others. An optimal model would portray the combined probability distribution of labels for each pixel based on the image, enabling the generation of various credible label configurations. Given that collecting more data does not reduce aleatoric uncertainty, it is imperative to explicitly model this uncertainty, especially in risk-sensitive applications such as medical image segmentation for diagnostic purposes. Within the domain of medical imaging, images often contain noise, and the boundaries between different tissue types might not always be clearly defined, resulting in differences of opinion even among experts. The capability of autonomously producing multiple credible hypotheses is invaluable, especially in fields such as radiotherapy, where decisions on which anatomical regions to target for invasive procedures must be made. Furthermore, offering confidence bounds in conjunction with tumour outlines would facilitate the incorporation of uncertainty in pivotal decision-making processes. [Monteiro et al., 2020].

In machine learning, there are various sources of uncertainty. Data collection can cause aleatoric uncertainty, which happens when not all necessary explanatory variables are included in the dataset [Papernot and McDaniel, 2018]. For example, if a dataset is used for segmenting medical images, such as MRI scans, and the annotations are given as bounding boxes instead of precise, pixel-level ground truths, aleatoric uncertainty is introduced. These bounding box annotations are helpful, but they do not precisely show the boundaries of areas like tumours, which can lead to variability in

the model’s predictions.

In this context, it is essential to devise a system to estimate aleatoric uncertainty, especially for medical professionals who depend on these models. This system would give physicians an understanding of the inherent uncertainty in the model’s predictions, which is especially important when the training data are not accurately labelled. This knowledge allows physicians to assess the dependability of the segmentation results and make more informed decisions, particularly when high accuracy is necessary. Furthermore, understanding the uncertainty can help them decide when extra diagnostic information or tests might be needed. Additionally, from a development standpoint, recognising areas of high aleatoric uncertainty can direct efforts towards improving data collection, thus increasing the model’s performance and trustworthiness in clinical settings.

In standard neural networks used for classification or regression, the final layer outputs a probability distribution over the set of classes or a distribution over a point prediction [Hüllermeier and Waegeman, 2021]. This captures aleatoric uncertainty, which is the uncertainty inherent in the data due to randomness or noise. Several methods have been proposed in the literature for estimating aleatoric uncertainty for DCNNs.

Depeweg et al. proposed an approach to estimating aleatoric uncertainty that involves a Bayesian framework combined with latent variables. In this method, the predictive distribution of a BNN for a target variable associated with a test data point is given by integrating over both the weights and latent variables [Depeweg et al., 2018]. The key idea is to decompose the total uncertainty into aleatoric and epistemic components. Aleatoric uncertainty, in this context, originates from the randomness of the latent variables and noise term, and it represents the uncertainty inherent in the

data that cannot be reduced by collecting more data. This method quantifies this uncertainty by considering the entropy or variance of the predictive distribution. The aleatoric uncertainty is then specifically isolated by computing the expectation of the entropy or variance under the model’s posterior distribution over the weights, effectively separating it from the epistemic uncertainty, which is associated with the model parameters.

In the method proposed by Kendall and Gal [Kendall and Gal, 2017], aleatoric uncertainty is estimated in the context of Bayesian deep learning for both regression and classification tasks. For regression, the approach involves heteroscedastic regression, where the observation noise varies with the input. This is achieved by modifying the loss function of a neural network to learn the variance of the noise as a function of the input data. In classification tasks, the method adapts by placing a Gaussian distribution over the logits (the vector of raw predictions from a neural network before the softmax layer) and learning the variance of this distribution [Kendall and Gal, 2017]. The key aspect of their method is the direct modeling of uncertainty within the data by learning the variance associated with the model’s output. In both regression and classification, the approach focuses on capturing aleatoric uncertainty by making the model’s output sensitive to the inherent noise in the observations, thus providing a tailored approach for handling uncertainty in different types of predictive tasks.

Jungo and Reyes used a method for estimating aleatoric uncertainty that involves designing a neural network that outputs both the mean and variance of its predictions. This approach is structured around the concept that the network, denoted as f , produces two outputs for each input x : the mean of the logits $\hat{x}$ and the variance σ^2 , which represents the aleatoric uncertainty. The logits are assumed to be perturbed with Gaussian noise, aligning with the concept of aleatoric uncertainty that accounts for the noise

inherent in the observations. The training of this network involves optimising a loss function that considers both the accuracy of the predictions (mean) and the confidence or uncertainty associated with these predictions (variance) [Jungo and Reyes, 2019]. Monte Carlo sampling is employed to effectively estimate the aleatoric uncertainty, generating multiple samples of the perturbed logits to approximate the expectation over the noise distribution. This model provides a direct and practical way to quantify the uncertainty associated with each prediction made by the neural network.

The three methods for estimating aleatoric uncertainty vary in their approaches and suitability. Depeweg’s method is a comprehensive Bayesian-based decomposition of uncertainty, Kendall’s method modifies neural network loss functions to learn data-inherent noise variance, and the method proposed in [Jungo and Reyes, 2019] is a straightforward, practical way of directly measuring uncertainty in predictions. The best method to use depends on factors such as the complexity of the model and the type of data.

3.3.3 Factors Influencing Uncertainty Map Interpretations

If we break down the sources of uncertainty affecting the predictions of a neural network, the uncertainty of the model includes variability in real-world situations, the architecture specification of the DCNN, the training setting and parameters of the DCNN, and errors caused by unknown data, while the uncertainty of the data pertains to error and noise in measurement systems. Although one can potentially minimise model uncertainty by refining the network architecture, optimising the learning process, or enhancing the training dataset, data uncertainties are inherent and cannot

be completely eliminated. It is essential for deep neural networks to handle these uncertainties proficiently, particularly when they are used in mission-critical or safety-sensitive real-world situations. [Gawlikowski et al., 2023].

Following the discussion on the sources of uncertainty, it is crucial to delve into the specific factors that contribute to these uncertainties in medical image segmentation.

Acquisition Methods

One such significant factor is the method of image acquisition. The choice of imaging modality, whether MRI, CT, or ultrasound, along with specific acquisition parameters, plays a crucial role in determining the quality and characteristics of captured images. For instance, MRI sequences with different contrast mechanisms can vary in their ability to capture certain pathologies, introducing aleatoric uncertainty due to inherent data variability. In Figure 3.1, we see a comparison of MS lesions as captured in T2 FLAIR versus T1-weighted MRI sequences. T2 FLAIR is adept at highlighting fluid-attenuated lesions, offering superior visibility of certain types of MS lesions, while T1-weighted sequences provide detailed anatomical information, which is crucial for understanding the lesion’s impact on surrounding tissues.

Another example is the variations caused by different scanners used to collect the data in the first place. Figure 3.2 depicts these differences. This variability is a critical factor in medical segmentation challenges, as highlighted in recent studies. Poor contrast or other limitations inherent to the acquisition process can lead to uncertainties that directly affect the segmentation outcome. In light of these challenges, it becomes imperative for segmentation models to not only adapt to the diverse range of data

presented by different imaging techniques but also to incorporate robust uncertainty estimation methods. These methods play a crucial role in capturing and quantifying the nuances of data variability, thereby enabling physicians to interpret the model outputs with a clearer understanding of their underlying uncertainties. This approach is essential for enhancing the trustworthiness and clinical utility of automated segmentation in medical diagnostics.

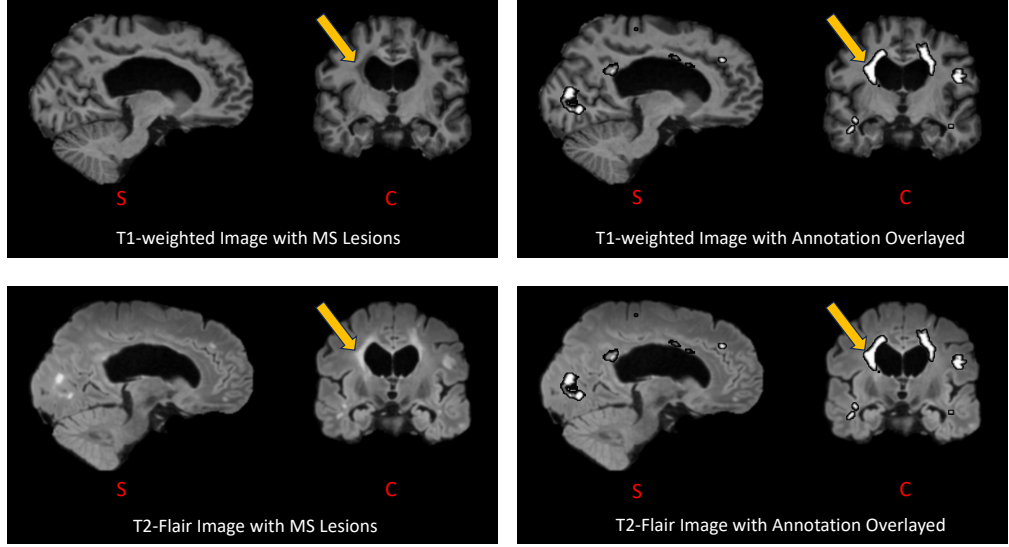


Figure 3.1: An example of MS Lesions in T2 FLAIR vs T1. Yellow arrows show a typical pathology in MS patients (white matter lesion) and how the lesion appears differently in T1 vs T2 FLAIR. T1 is excellent in showing grey and white matter, while T2 FLAIR is suitable for suppressing fluid signals and depicting tissue alteration such as white matter lesions hence the hyperintensity segments.

Image Quality

The quality of an image is determined by factors such as resolution, SNR, and the presence of artefacts which can directly influence the performance and reliability of segmentation algorithms. High-quality images may reduce aleatoric uncertainty but could introduce epistemic uncertainty if the model is not trained on such data:

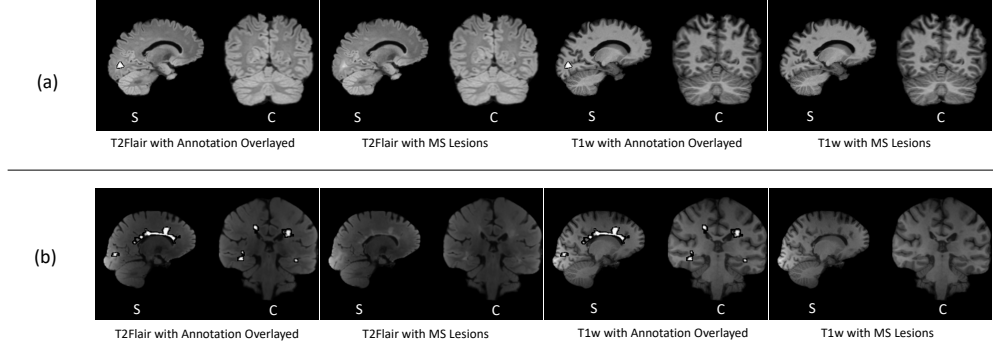


Figure 3.2: (a) a patient with MS MRI images with T1w and T2Flair sequences acquired on a 3T Achieva MR scanner (Philips Medical, Best, The Netherlands) using a 32 channel head coil. (b) a patient with MS MRI images with T1-weighted and T2 FLAIR sequences acquired on a 3T MR750 scanner (GE Healthcare, Milwaukee, USA) using a 32 channel head coil

The quality of medical images is largely dependent on their resolution. High-resolution images offer more precise details, allowing for more accurate segmentation of fine structures. On the other hand, low-resolution images may cause uncertainty in segmentation due to the lack of detail, potentially missing small but clinically relevant features such as tiny lesions or subtle anatomical differences. In [Yap et al., 2017] a comparative analysis of two distinct breast lesion datasets (Dataset A and Dataset B) is provided, highlighting variations in speckle noise, overall image quality, and lesion depiction. In this comparison, Dataset B stands out for its reduced speckle noise, attributed to the advanced ultrasound acquisition system employed. This system’s capability is further evident in the superior image quality of Dataset B, where the resolution surpasses that of the older technology used in Dataset A. Such enhanced resolution in Dataset B leads to more discernible anatomical structures, including ribs, pectoral muscles, and parenchymal tissue. Lesions in Dataset B are depicted with greater clarity, allowing for the visualisation of intricate internal structures within fibroadenoma lesions. Figure 3.3 shows some of these high resolution images in Dataset B. This level of detail in Dataset B is crucial, as

high-resolution images provide more precise information, enabling more accurate segmentation of fine structures. Conversely, the lower resolution in Dataset A might introduce uncertainty in segmentation, potentially overlooking small yet clinically significant features. For a comprehensive comparison, including images from both datasets, readers are referred to the original paper as Dataset A is not publicly accessible [Yap et al., 2017]. The contrast between the datasets underscores the impact of resolution on the reliability of medical image segmentation.

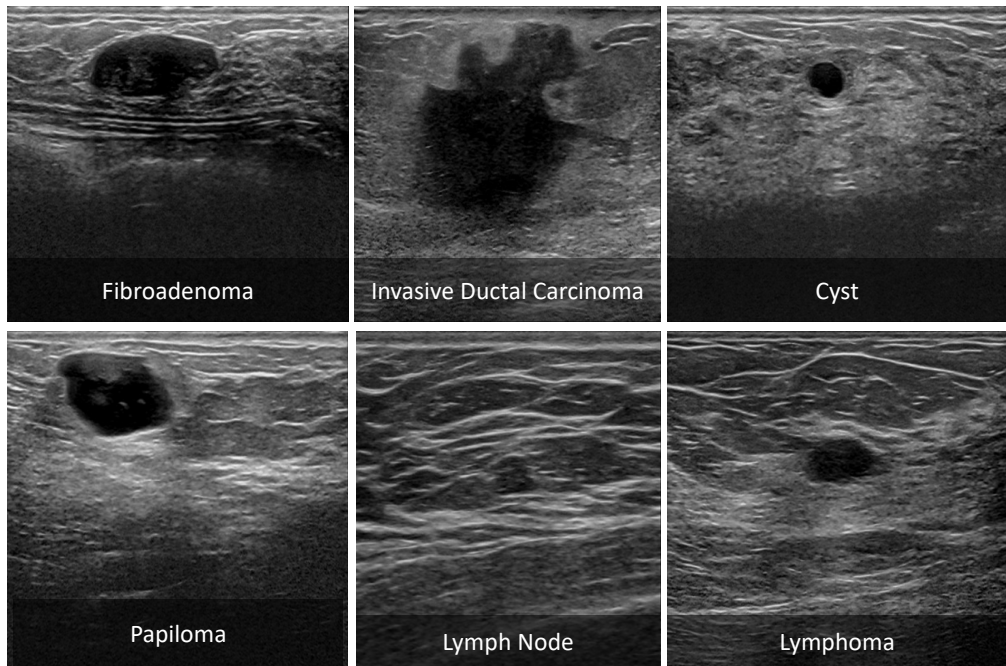


Figure 3.3: Examples of high-resolution images in Breast Ultrasound Lesions Dataset (Dataset B) including Fibroadenoma, Invasive Ductal Carcinoma, Cyst, Papiloma, Lymph Node and Lymphoma.

Signal-to-noise ratio is a key measure of image quality that reflects the clarity of the image. High SNR images have less background noise, allowing for clearer distinction between different tissues and more accurate segmentation [Diaz et al., 2011]. Conversely, images with low SNR are prone to aleatoric uncertainty due to the increased noise level, which can obscure critical details and lead to inaccurate segmentations. The Multi-

modal Brain Tumour Segmentation challenge (BraTS) 2018 dataset is a great example to discuss here. BraTS 2018 provides an extensive dataset of multimodal MRI scans for brain tumour segmentation, with a focus on clinically-acquired pre operative scans of glioblastoma and lower grade glioma [Menze et al., 2014, Bakas et al., 2017b, 2018].

Figure 3.4 shows several samples from the BraTS 2018 dataset, depicting the noise in images causing low SNR. Feng et al. showed that the use of preprocessing techniques, such as denoising methods [Feng et al., 2020], could significantly enhance the SNR of the images. This enhancement in image quality through preprocessing has implications for the nature of uncertainty in segmentation models. Models trained on images with low SNR must contend with higher levels of aleatoric uncertainty, as noise can obscure critical image details. In contrast, models trained on preprocessed images with higher SNR may face reduced aleatoric uncertainty but could encounter different challenges, such as epistemic uncertainty, if the preprocessing introduces new patterns or alters image characteristics in ways not represented in the training data. Thus, the choice of preprocessing techniques and the quality of training data are crucial in determining the type and extent of uncertainty that segmentation models must address.

Artefacts in medical imaging, arising from various sources such as equipment limitations, patient movement, or magnetic field variations, can significantly degrade image quality. These artefacts introduce errors and distortions that are not part of the actual anatomy, challenging the segmentation process. Effective handling of artefacts is crucial, as their presence can lead to both aleatoric and epistemic uncertainties in segmentation outcomes. In the realm of skin lesion imaging, particularly with the use of dermoscopy, various artefacts introduce significant challenges and uncer-

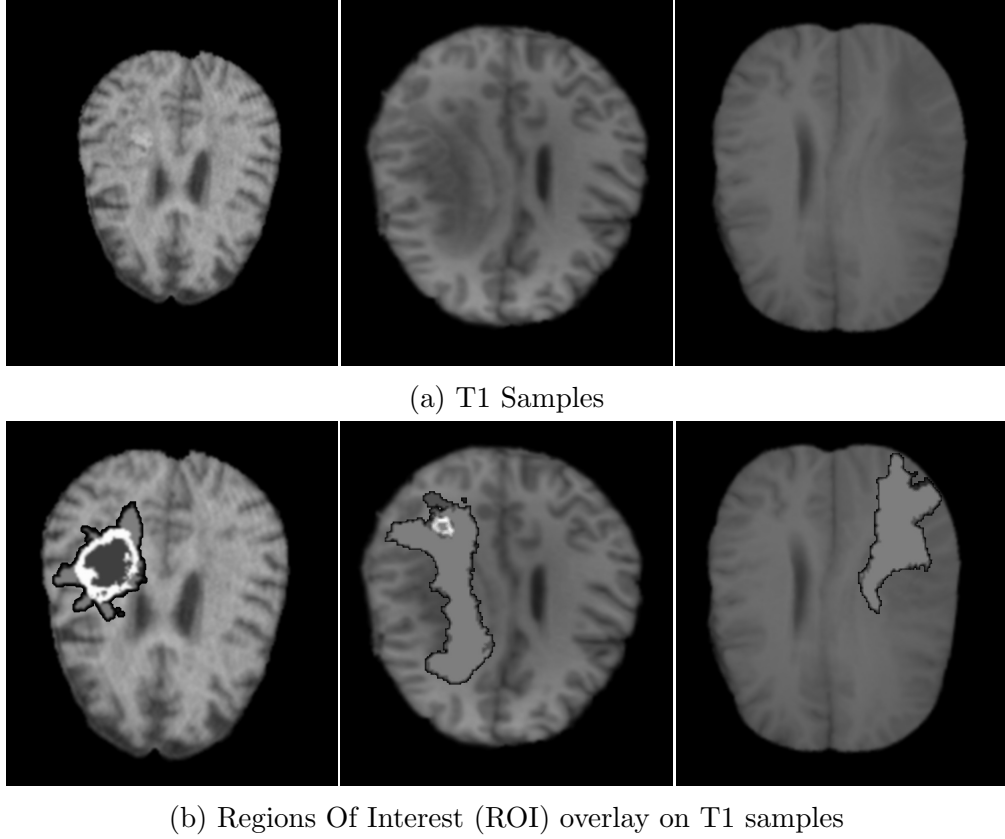


Figure 3.4: Brain tumour images from the BraTS 2018 dataset depicting the noise in images causing low SNR. High SNR images offer clearer differentiation of tumour boundaries, whereas low SNR images introduce noise, complicating accurate segmentation.

tainties in image interpretation. Common artefacts in dermoscopic images, such as dark corners, hair, colour charts, rulers, measurement markers, and ink markers, [Mahbod et al., 2019] each contribute to aleatoric uncertainty. These elements can obscure the boundaries of the lesion, cover important areas, or introduce misleading features, complicating the task of accurately identifying and segmenting the lesions. Furthermore, the use of interface fluids in contact dermoscopy, while enhancing visualisation, can introduce artefacts like air bubbles [Chen et al., 2014, Benvenuto-Andrade et al., 2007]. These bubbles may appear as bright spots, potentially mimicking or hiding critical lesion characteristics, thereby adding to the uncertainty in distinguishing true pathological features from artefacts. This interplay

of artefacts not only challenges the accuracy of automated analysis but also underscores the importance of sophisticated image processing techniques and uncertainty-aware models in dermatology. Such models must be adept at differentiating between genuine lesion attributes and misleading artefacts to provide reliable and precise diagnostic information, crucial for effective treatment planning and patient care. Figure 3.5 shows some of these artefacts that appear in the public dermoscopy dataset from the ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection challenge (ISIC2018) dataset [Tschandl et al., 2018].

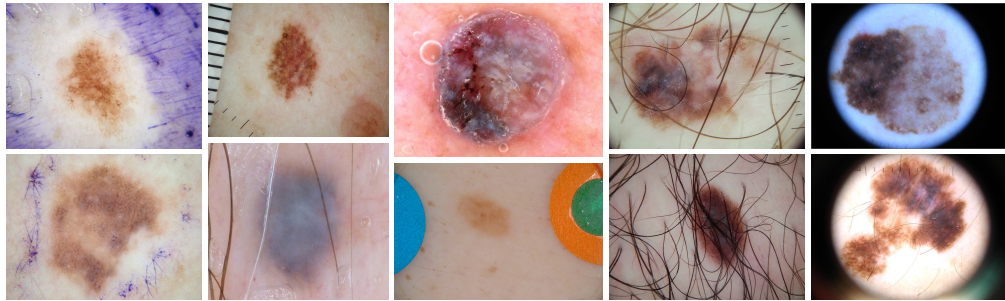


Figure 3.5: Examples of common artefacts in dermoscopic skin lesion images, including dark corners, hair, color charts, rulers, measurement markers, and ink markers, as well as the presence of air bubbles due to interface fluids. These artefacts introduce aleatoric uncertainty by obscuring or altering lesion characteristics, complicating accurate segmentation and analysis. Understanding and mitigating these artefacts is crucial for reducing uncertainty and enhancing the reliability of diagnostic interpretations.

Out of Distribution Data

In the realm of medical imaging, the effective deployment of DCNNs hinges not only on their accuracy but also on their ability to handle Out-Of-Distribution (OOD) data. OOD samples, which include rare diseases or atypical anatomical variations, can significantly impact the performance of DCNNs. These samples are noticeably different from those in the training set, leading to increased epistemic uncertainty. This uncertainty, if not properly accounted for, can undermine the reliability of the model.

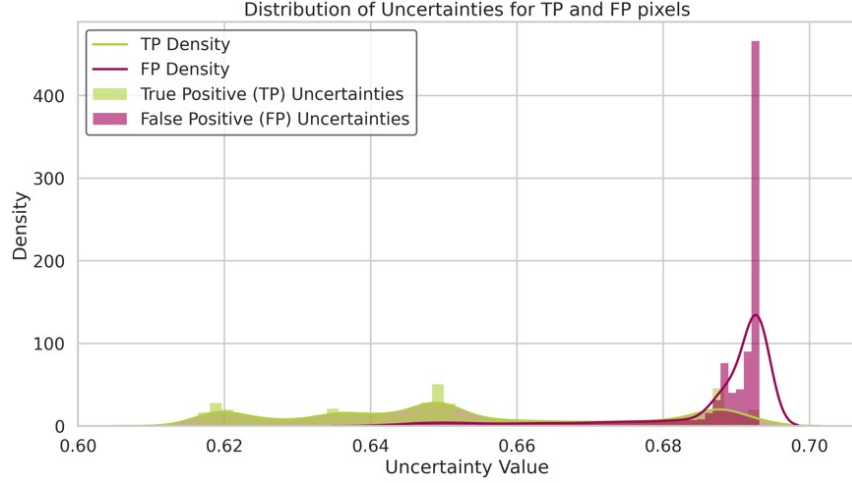
A crucial aspect of managing this challenge is the generation and interpretation of uncertainty maps at the time of inference. When a DCNN encounters OOD samples, the uncertainty map generated should accurately reflect the increase in uncertainty. This transparency is vital for medical professionals who rely on these models for diagnostic support. By examining both the segmentation results and their corresponding uncertainty estimations, experts can gain insights into the confidence level of the model predictions. This practice is essential to maintain consistency in interpretation and decision making.

However, as noted in the study by Linmans et al., DCNNs often exhibit overconfidence in their predictions, even when dealing with OOD data [Linmans et al., 2023]. This is where the prediction is incorrect but the model confidence is very high such as incorrectly segmenting a part of a tumour (foreground) as background. This overconfidence can mask the true level of uncertainty, leading to potentially misleading interpretations. To address this, uncertainty estimation provides a more nuanced view of the predictive distribution, thereby offering a more accurate representation of uncertainty. By incorporating these methods, DCNNs can produce uncertainty maps that more accurately reflect confidence levels in various scenarios, including encounters with OOD samples.

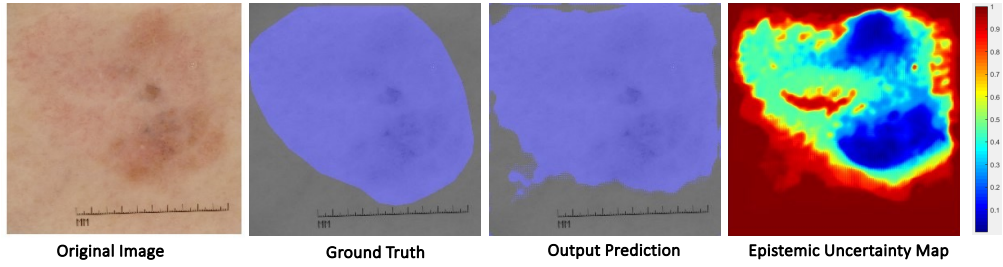
In addition to the broader challenges of OOD data, specific lesion characteristics such as size, shape, and contrast against surrounding tissues play a pivotal role in influencing uncertainty in medical imaging. For example, when models trained on a large dataset of small lesions, encompassing both well-defined and poorly contrasted boundaries, are exposed to significantly larger lesions with unusual textures at testing, the level of uncertainty can be noticeably different. In such cases, the model’s experience with predominantly smaller lesions may lead to increased epistemic uncertainty when

encountering these larger, texturally-unique lesions. This is because the model lacks prior exposure to such characteristics, making it less certain about its predictions. The uncertainty maps generated in these scenarios are crucial, as they visually represent areas where the model’s confidence is low.

For instance, a model accustomed to small lesions might display high uncertainty levels when analysing larger lesions with atypical textures, indicating a potential decrease in prediction reliability. To illustrate this, Figure 3.6 provides a comprehensive visual representation which includes the original image of a large lesion, which is distinct from the smaller lesions the model was trained on, alongside the ground truth annotation and the model prediction. It also presents the uncertainty map generated by the model for this particular case. The map vividly highlights areas where the model confidence diverges significantly from the ground truth. Additionally, the figure includes the uncertainty distribution of the pixels, distinguishing between the False Positive (FP) and True Positive (TP) pixels (TP and FP were chosen to focus on assessing the ability of epistemic uncertainty map in distinguishing between the pixels that were incorrectly labelled as lesion versus the pixels that the model correctly identified as lesion). This breakdown depicts how the model epistemic uncertainty varies across different areas of the lesion, reflecting the challenges it faces when dealing with larger lesions that exhibit atypical textures compared to its training data. Such visualisations are essential in understanding the model performance limits and guiding further improvements in training for enhanced accuracy and reliability in diverse clinical scenarios.



(a) Epistemic Uncertainty Distribution of TP and FP Pixels



(b) Model Output Result for a Large Lesion

Figure 3.6: (a) Distribution of uncertainty in TP and FP pixels (generated from calculating epistemic uncertainty map at inference time), highlighting the model’s confidence variability in lesion segmentation, showing generally high epistemic uncertainty for both TP and FP pixels. (b) Visualisation of the model’s output for a large lesion, including the original image, ground truth, prediction, and uncertainty map, demonstrating the model’s response to lesion sizes beyond its training scope.

The challenges posed by OOD data in medical diagnostics are not limited to image segmentation alone, but extend to classification tasks as well. Despite the remarkable success of DCNNs in diagnosing various diseases, they exhibit a critical vulnerability: overconfidence in incorrect predictions when encountering images outside their training distribution. This phenomenon is particularly concerning in healthcare applications, where misclassification can have serious consequences [Zaid et al., 2022]. In the study, Zaid et al. discuss that even minor perturbations or noise in an image can lead

to high-confidence but incorrect predictions by the network. More alarmingly, images completely unrelated to the trained classes, such as those of a dog, can be misclassified as a skin cancer image by advanced CNN architectures like DenseNet [Huang et al., 2017], MobileNet [Howard et al., 2017], ResNet [He et al., 2016a], and VGGNet [Simonyan and Zisserman, 2014]. This issue underscores the necessity for DCNNs not only to accurately classify known conditions but also to reliably flag new or unfamiliar disease presentations [Zaid et al., 2022].

Translating these insights into the realm of medical image *segmentation*, the implications are profound. Just as misclassifications in disease diagnosis can lead to inappropriate treatment decisions, inaccuracies in segmentation due to OOD data can lead to incorrect assessments of the extent, progression, or even the nature of the disease itself. Therefore, segmentation models must be equipped to handle OOD samples effectively, either by accurately segmenting them or flagging them for further expert review.

It is also crucial to recognise the challenges involved in collecting and precisely annotating out-of-distribution (OOD) samples, particularly in the medical field where the spectrum of potential OOD scenarios is vast and unforeseeable [Zaid et al., 2022]. Figure 3.7 shows the distribution of in-distribution (ID) and OOD in the Breast Ultrasound Lesions dataset (BUS17) [Yap et al., 2017]. This difficulty adds another layer of complexity to the development of robust DCNNs that can effectively manage such data. In addition, it introduces added uncertainty, particularly in the face of the complexities associated with challenging annotations.

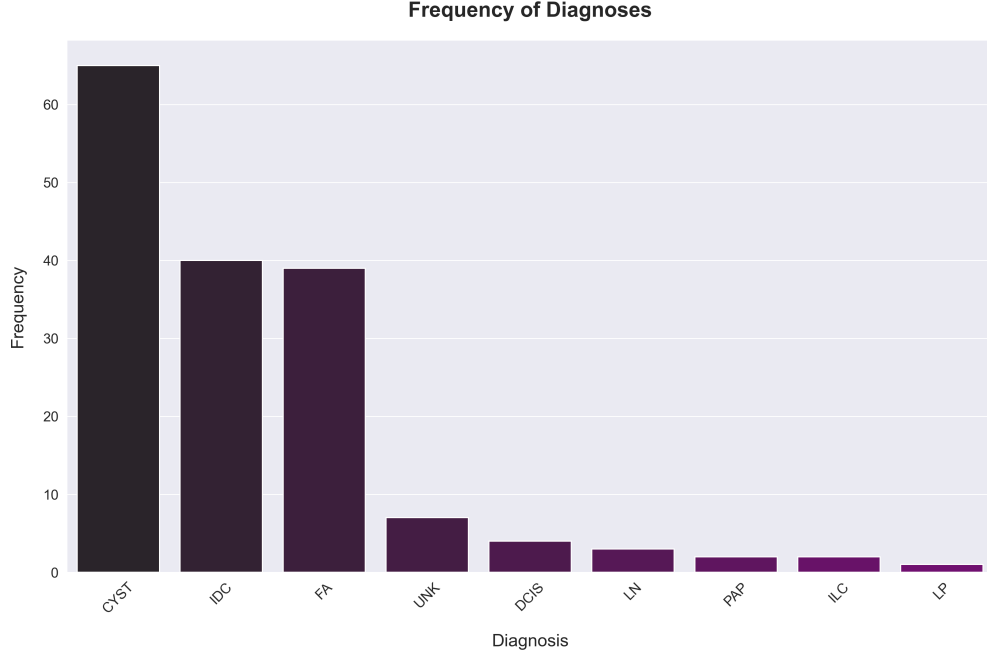


Figure 3.7: Histogram plot of the breast cancer lesion dataset categorised by lesion types. The tail of this histogram depicts the OOD samples including Unknown (UNK), Ductal Carcinoma In Situ (DCIS), Lymph Node (LN), Papiloma (PAP), Invasive Lobular Carcinoma (ILC), Lymphoma (LP) lesions.

Addressing OOD data is critical in image segmentation. The development of DCNNs that are not only accurate in familiar scenarios, but also capable of indicating their limitations when encountering novel or rare cases, is essential. This approach is vital to ensure the safety and reliability of Artificial Intelligence (AI)-assisted medical diagnostics in both classification and segmentation tasks. Furthermore, the ability of Deep Convolutional Neural Networks (DCNN)s to produce accurate and interpretable uncertainty maps in the presence of Out-Of-Distribution (OOD) data is paramount in medical imaging applications. This capability is essential to maintain consistent interpretations by medical experts and to improve the trustworthiness and reliability of AI-assisted diagnostic systems. As the field continues to evolve, a sustained emphasis on improving uncertainty estimation methods will be key to overcoming the challenges presented by

OOD data.

Annotation Variability

In supervised learning, particularly in medical image segmentation, the variability in expert annotations plays a critical role in shaping the model’s uncertainty. As highlighted by Jungo et al., inter-observer variability introduces aleatoric uncertainty due to differences in expert opinions, which is inherent in medical applications. This variability, if not properly accounted for, can lead to unreliable uncertainty estimations [Jungo et al., 2018a]. The study demonstrates that different fusion methods used to combine expert annotations, such as majority voting [Rogova, 2008], significantly impact the model’s ability to estimate uncertainty accurately. Particularly, avoiding label fusion (i. e. combining several different labels of the same entity into a single discrete label [Robitaille et al., 2012]) in training data can lead to more reliable uncertainty estimations that better reflect expert disagreement. This suggests that the choice of annotation and labeling strategies is essential not only for model accuracy but also for the effective quantification and interpretation of uncertainty in predictions, underscoring the need for careful consideration of inter-rater variability in model development and validation processes.

Model Complexity, Training Data Diversity and Volume

The complexity of a neural network, characterised by the number of layers and parameters, can significantly impact its uncertainty handling capabilities. Overly complex models have a higher risk of overfitting to the training data. Although this might reduce aleatoric uncertainty by closely modelling the noise and nuances in the training set, it can increase epis-

temic uncertainty. This increase is due to the model’s reduced ability to generalise to new, unseen data. In essence, a complex model may perform exceptionally well on training data, but fail to maintain this performance on new data, reflecting high epistemic uncertainty.

Furthermore, the diversity and volume of training data are equally crucial. A dataset that is limited or not representative of the segmentation task at hand can lead to increased epistemic uncertainty. In contrast, a large and diverse dataset can significantly reduce epistemic uncertainty as it provides the model with a more comprehensive view of the problem space, allowing for better generalisation.

The use of pre-trained models [Kalinin et al., 2020, Stuckner et al., 2022, Stawiaski, 2018] in machine learning tasks can have mixed implications for epistemic uncertainty. These models can be advantageous, particularly in scenarios where the training data available for a specific task are limited. Using a well-established foundational knowledge base, pre-trained models can effectively diminish epistemic uncertainty. However, challenges arise when the domain from which the pre-trained model was initially developed significantly diverges from the target application domain [Mehta et al., 2022, Oquab et al., 2014]. Such discrepancies can inadvertently lead to an increase in epistemic uncertainty due to domain misalignment. In these instances, the application of domain adaptation techniques [Bousmalis et al., 2016] becomes essential. These techniques are instrumental in adjusting pre-trained models to more closely align with the unique characteristics and requirements of the new task, thus mitigating the risks of increased uncertainty.

Achieving an optimal balance between the complexity of a neural network model and the characteristics of the training data is crucial for uncer-

tainty management. When a model is excessively complex relative to the available data, it tends to overfit, which results in poor generalisation and consequently increased epistemic uncertainty. Conversely, a model that is overly simplistic may fail to capture intricate patterns present in the data, leading to increased aleatoric uncertainty. Additionally, the nature of the training data plays a central role. It is essential for the training dataset to be not only diverse and large enough to effectively train complex models, but also to be closely aligned with the specific domain of the task at hand. This alignment ensures that the model learns relevant patterns and generalises well to new, unseen data.

3.4 Methodology

As we discussed in section 3.3.2, understanding and addressing uncertainty is crucial, particularly distinguishing between aleatoric and epistemic uncertainty. We mentioned that aleatoric uncertainty, inherent and irreducible, arises from the stochastic nature of the data itself. It is exemplified by the intrinsic randomness in the relationship between input variables and outcomes, as captured by the conditional probability distribution. This type of uncertainty remains constant even with complete knowledge of the underlying probability distributions and cannot be mitigated by additional data or information. On the other hand, we learned that epistemic uncertainty stems from a lack of knowledge or information, particularly about the perfect model or predictor. This uncertainty is reducible; it can be diminished by acquiring more data or enhancing our understanding of the model parameters [Hüllermeier and Waegeman, 2021].

However, it is important to also note that the distinction between these

two types of uncertainty is not static but context-dependent. For instance, introducing additional features or altering the learning environment can transform what was initially considered aleatoric uncertainty into epistemic uncertainty. This transformation is evident in scenarios where, for example, two class distributions that overlap in a low-dimensional feature space become separable when the data are embedded in a higher-dimensional space. Such changes highlight how the introduction of new features or adjustments in the model's complexity can shift the nature of the uncertainty from being inherent in the data (aleatoric) to being a function of our knowledge and model sophistication (epistemic) [Hüllermeier and Waegeman, 2021].

In a medical image segmentation context, the transition from aleatoric to epistemic uncertainty can be illustrated by the shift from low-quality to high-quality images. Initially, segmentation models face aleatoric uncertainty due to the inherent noise and imprecision in low-quality images, particularly around tumour/lesion boundaries. This uncertainty is intrinsic to data quality and cannot be resolved by improving the model alone. However, with the introduction of high-quality imaging, the nature of uncertainty changes. The enhanced clarity and detail of the images reduce the inherent data noise, but could introduce epistemic uncertainty due to the model's limitations. The existing model, trained on lower-quality images, may not be sophisticated enough to interpret finer details in high-quality images. This scenario depicted in Figure 3.8, shows how changes in data quality can transform inherent data uncertainty (aleatoric) into uncertainty stemming from model limitations (epistemic). This dynamic nature of uncertainty in medical image segmentation underscores the importance of a flexible and informed approach in modelling and decision-making processes, where the categorisation and handling of uncertainty must adapt to the evolving context and available information.

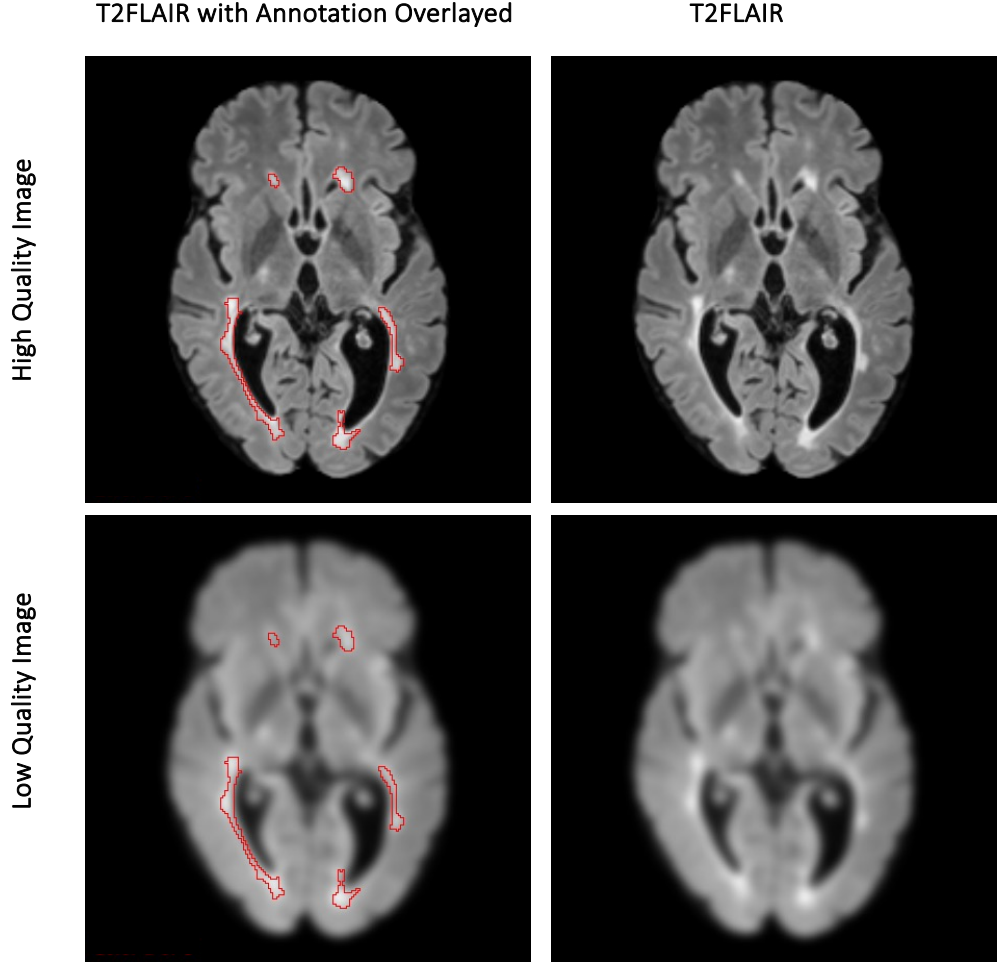


Figure 3.8: How aleatoric uncertainty can become epistemic uncertainty in when the model trained on low quality data encounters high quality data with fine details that the model has not seen before. The lesions are contoured with red lines.

Recent studies, including those by Jungo et al., have underscored a critical challenge in the deployment of deep learning systems in clinical settings: while accuracy has seen notable improvements, robustness remains a concern, particularly at the subject level. These studies reveal that common voxel-wise uncertainty measures, though well-calibrated at the dataset level, often exhibit miscalibration when it comes to individual subjects [Jungo and Reyes, 2019]. This inconsistency leads to considerable concerns regarding the reliability of uncertainty estimates in clinical settings, where the accuracy at the individual patient level is of high importance.

In response to this challenge, we propose leveraging uncertainty as an attention mechanism within our model. This approach aims to enhance subject-level calibration, ensuring that uncertainty measures are not only globally accurate, but also finely adjusted to individual patient data. Furthermore, the critical nature of medical imaging, a field characterised by high stakes and the need for utmost precision, necessitates a bespoke approach to model design. Recognising this, we emphasise the importance of tailoring the loss function for medical image segmentation.

Our proposed uncertainty-aware loss functions are designed to integrate uncertainty information directly into the training process, guiding the model to focus on areas of high uncertainty. This approach may need to be tailored for each medical imaging scenario or disease, as it is not a universal solution but a targeted strategy for the specific challenges in medical imaging tasks. To thoroughly evaluate the effectiveness and potential limitations of this approach, we plan to conduct extensive assessments across various datasets. Through these evaluations, we aim to not only validate the utility of our proposed method in enhancing subject-level calibration but also to explore its broader applicability and identify any areas for refinement. This testing is crucial to ensure that our solution is robust, reliable, and suitable for the delicate and critical context of medical imaging.

The overview of the the proposed architecture is shown in Figure 3.9. The proposed model includes a Unet [Ronneberger et al., 2015] architecture, with tailored designed loss function. The first proposed tailored loss function includes the attention map block calculated by the epistemic uncertainty applied to cross entropy loss in the form of weight map. The second proposed tailored loss function is similar to the the first loss, except the attention map block is calculated by the error of the model. Finally, the third set of proposed tailored loss functions are the combination of aleatoric

loss as regularisation term with one of the error-based or uncertainty-based attention maps. The details of each of the proposed tailored loss functions are discussed in the next subsections.

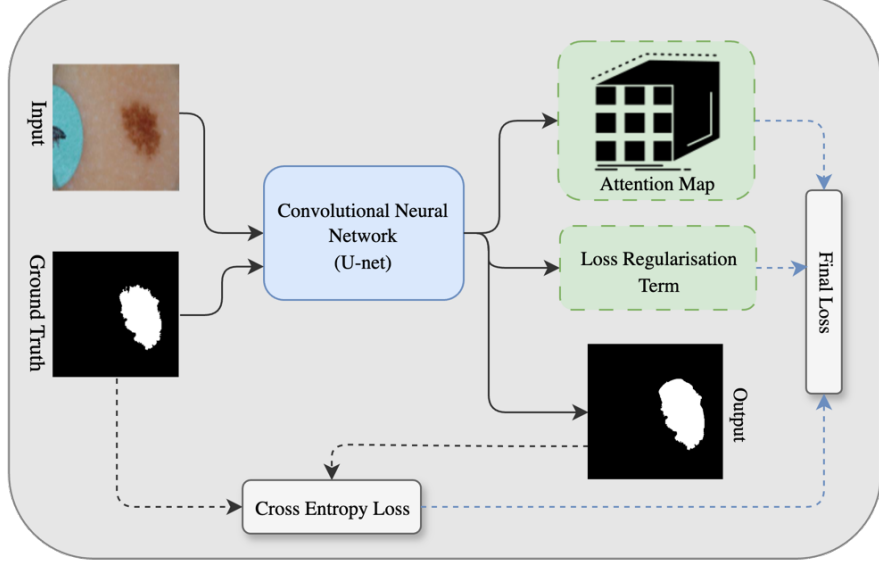


Figure 3.9: The overall architecture of the proposed models

3.4.1 Encoder-decoder DCNN for Image Segmentation

The basic architecture of our image segmentation network is a well-established DCNN method (i.e. Unet [Ronneberger et al., 2015]). The Unet architecture consists of a multilayer encoding path to capture context and a symmetric decoding path that enables the production of high-resolution segmentation results. In the encoder part, each layer consists of two 3×3 convolutions, a ReLU [Ronneberger et al., 2015] and a 2×2 max pooling of stride 2. In the decoder part, each layer consists of upsampling of the feature map, a 2×2 up-convolution which halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the encoding path, and two 3×3 convolutions, each with a ReLU. We

also add dropout [Srivastava et al., 2014] and batch normalisation [Ioffe and Szegedy, 2015] after each convolutional layer. A 1×1 convolution is used in the final layer to generate the segmentation output. The cross entropy loss function is used for optimisation of network parameters, as expressed in 3.4.

$$\mathcal{L}_{ce}(p(x), g(x)) = - \sum_{c=0}^{c-1} g_c \log(p(y = c|x)) \quad (3.4)$$

where g_c is the ground truth label of class $c \in C$ and $p(y = c|x)$ is the model’s estimated probability for class c given input x .

3.4.2 Uncertainty Attention Map

Using the pixel-wise uncertainty map as an attention mechanism in the loss function, the model primarily targets the reduction of epistemic uncertainty, with potential indirect benefits to how the model handles aleatoric uncertainty. This research examines the possibility of this approach resulting in a more adaptable learning process. We assess whether this could lead to a more robust model and more reliable uncertainty interpretation.

As mentioned in section 3.3.2, one of the most popular stochastic regularisation techniques for deep learning models is dropout [Gal, 2016]. Dropout is used to avoid overfitting the network, which can be interpreted as a form of approximate inference in a probabilistic interpretation of the model [Gal and Ghahramani, 2016, Srivastava et al., 2014]. This allows us to model uncertainty using dropout. Predictive entropy can express a measure of ignorance about the actual structure of the model that captures the amount of information in the predictive distribution [Gal, 2016]. The predictive entropy $H[y|x, \mathcal{D}]$ (equation 3.1 in Section 3.3.2) presents the epistemic uncertainty where the values range between 0 and 1. The higher the value

of $H[y|x, \mathcal{D}]$, the more uncertain the prediction will be. For detailed calculation of the epistemic uncertainty the reader is referred to Section 3.3.2.

The obtained epistemic uncertainty maps are normally used at inference time in the literature. In our case, the estimated pixel-wise uncertainty map is used in the loss function as an attention mechanism. Intuitively, a higher weight is assigned to the pixel location with higher uncertainty, leading to a larger contribution to the loss calculation to guide network optimisation. This forces the network to focus on learning from image regions with higher uncertainties. We propose to calculate the weight map as follows:

$$M_{unc}(x) = \alpha^{(1-H[y|x, \mathcal{D}])^\beta} \quad (3.5)$$

where $\alpha \in \mathbb{R}_+^*$ denotes the minimum weight value for a pixel and $\beta \in [1, 3]$ determines the behaviour of the nonlinear mapping for weight calculation. Here, we set the value of α to be 0.1, which prevents the network from assigning very small or zero weights to image regions with low uncertainty values. The values set for parameters α and β are selected experimentally. Figure 3.10 illustrates different uncertainty weight functions based on different settings of β .

By combining the uncertainty map with cross entropy loss, the final proposed loss function is expressed as

$$\mathcal{L}_{UCF} = \sum_{x \in \mathcal{X}} [M_{unc}(x) \times \mathcal{L}_{ce}(p(x), g(x))] \quad (3.6)$$

where x represents the individual pixel in an image X . The loss function in 3.6 is minimised using the U-net model introduced in Section 3.4.1. We call this method the *Uncertainty-Calibrated loss Function (UCF)* in our experiments.

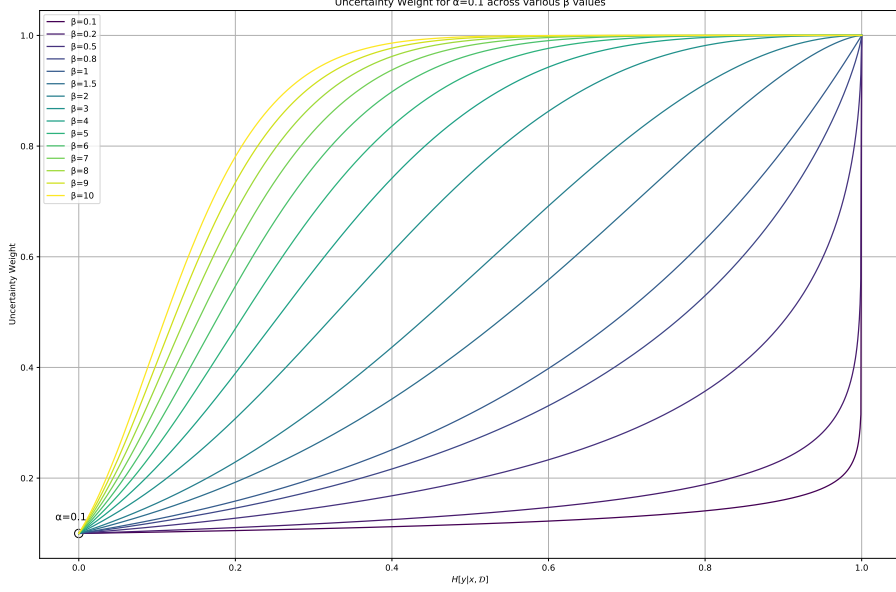


Figure 3.10: Variation of the Uncertainty Weight Function for Different Values of β : These plots illustrate how the uncertainty weight function changes with the entropy $H[y|x, \mathcal{D}]$ for a range of β values, given fixed α value of 0.1. Each line represents a distinct β value, demonstrating the influence of β on the weight function response to entropy. The starting point of each line, marked by a circle, corresponds to the value of α , highlighting the initial influence of α on the uncertainty weight.

3.4.3 Error-Weighted Loss

In order to assess the impact of uncertainty-aware weight map in the training process, we proposed a loss function with error-based weight map to be compared to the uncertainty-aware loss function.

Recent studies have utilised prediction error to develop error-weighted loss functions, significantly enhancing training effectiveness. An example of these functions is Focal Loss [Lin et al., 2017], which modifies cross-entropy loss to prioritise learning from challenging examples, particularly in object detection. Another important development is Focal Tversky Loss [Abraham and Khan, 2019], emphasising the need to focus on hard examples, especially in imbalanced datasets and diverse data sizes.

In this study, we introduce an error-based weight map inspired from these

approaches. This map adjusts the loss for each pixel based on its predictive error, encouraging the model to learn more from areas where predictions are less accurate. The weight for a pixel is related to its prediction error, thus directing the network’s learning focus towards regions with higher predictive challenges.

For a balanced comparison with our uncertainty-weighted loss, we employ a similar non-linear function:

$$M_{\text{err}}(x) = \lambda^{(1-\text{Error}(x))^\gamma} \quad (3.7)$$

Here, $\text{Error}(x)$ denotes the difference between the model’s estimated probability for the correct label of pixel x and the actual label. The hyperparameters λ and γ mirror α and β in our uncertainty map equation 4.1.

The resulting error-weighted loss function, used in our experiments and referred to as *Error-focused Loss Function (ELF)*, is defined as:

$$\mathcal{L}_{\text{ELF}} = \sum_{x \in \mathcal{X}} [M_{\text{err}}(x) \times \mathcal{L}_{\text{ce}}(p(x), g(x))] \quad (3.8)$$

It is important to note that difference between the use of ELF and UCF is the approach used to guide the model learning. ELFs target areas where the model’s predictions are inaccurate, prioritising hard-to-classify examples. In contrast, UCFs emphasise regions where the model is unsure of its predictions, thus aiming to improve the model’s confidence and reliability across the dataset.

3.4.4 Aleatoric Loss

In our research, we employed a loss function inspired by Jungo and Reyes method [Jungo and Reyes, 2019], specifically designed to encapsulate aleatoric uncertainty for a thorough evaluation and comparison with our proposed loss functions. This loss function is derived from a neural network structure, f , that yields two outputs for each input x : the mean of the logits $\hat{x}$ and the variance σ^2 [Jungo and Reyes, 2019]. $\hat{x}$ is the network’s prediction while σ^2 quantifies the aleatoric uncertainty, reflecting the inherent noise within the data. The design of this loss function allows for the integration of the aleatoric uncertainty associated with the predictions, treating the variance of the logits as an essential component of the loss calculation. This is done by incorporating Gaussian noise into the logits, and through Monte Carlo sampling.

$$\mathcal{L}_{aleatoric} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{1}{S} \sum_{s=1}^S P(y_i | \hat{x}_{i,s}, \sigma_{i,s}^2) \right) \quad (3.9)$$

N denotes the total number of data points under consideration. The term S represents the number of Monte Carlo samples, which are used to approximate the expectation over the noise distribution. For each data point indexed by i , y_i indicates the true label, providing the ground truth against which predictions are compared. The term $\hat{x}_{i,s}$ is used to represent the s -th sampled logits corresponding to the i -th data point. The variable $\sigma_{i,s}^2$ quantifies the variance of the logits for the i -th data point and s -th sample, reflecting the model’s uncertainty about its prediction. Lastly, $P(y_i | \hat{x}_{i,s}, \sigma_{i,s}^2)$ denotes the calculated probability of the true label given the sampled logits.

3.4.5 Combination Loss Functions

The Aleatoric Loss (AL) is combined with the other two proposed loss functions in order to offer a more extensive assessment and comprehension of the proposed models and their impact on the interpretation of the uncertainty map. Therefore, the Combination loss functions are as follows:

$$\mathcal{L}_{\text{UCF+AL}} = \mathcal{L}_{\text{UCF}} + \lambda \mathcal{L}_{\text{Aleatoric}} \quad (3.10)$$

where λ is for balancing the UCF and Aleatoric terms.

$$\mathcal{L}_{\text{ELF+AL}} = \mathcal{L}_{\text{ELF}} + \lambda \mathcal{L}_{\text{Aleatoric}} \quad (3.11)$$

where λ is for balancing the ELF and Aleatoric terms.

3.5 Experiments

3.5.1 Datasets

To investigate the understanding of the uncertainty map, the influencing factors discussed in Section 3.3.3 and the proposed methods, three challenging and varied datasets were chosen for the experiments: 1) ISIC2018, Skin Lesion Dataset, 2) BraTS, Brain Tumour Dataset and 3) BUS17, Breast Lesion Dataset. These datasets were specifically chosen due to their distinct characteristics and the diverse challenges they present, which are essential for a comprehensive evaluation of uncertainty in medical image segmentation.

ISIC2018, Skin Lesion Dataset

The first dataset is a public dermatoscopy dataset from the ISIC2018 challenge [Tschandl et al., 2018]. This dataset consists of 2,594 RGB dermoscopic images of different types of skin lesions with corresponding lesion segmentation masks.

There are some interesting points about this dataset that models may find quite challenging. First, the dermatoscopy dataset contains images of poor contrast, for some, the skin lesions and background may even appear identical. This is despite the fact that during the quality review process for the dataset workup, manual histogram corrections were applied to enhance the contrast and colour reproduction of the images [Tschandl et al., 2018]. The second point is the existence of varied textures in the background and foreground. Third, there are various lesion scales in this dataset. For some subjects in this dataset, the lesion segmentation may not even appear trivial to a trained human expert [Ibtehaz and Rahman, 2020]. Then there is the presence of various artefacts such as hair, colour charts, rulers, dark corners, measurement markers, and ink markers, as well as the presence of air bubbles due to interface fluids. Such challenges are not limited to this dermatoscopy dataset, but similar situations can be seen in other types of medical image segmentation datasets. Therefore, the dermatoscopy dataset is useful as a diverse representative set of samples that introduce realistic challenges to medical image segmentation models. In our experiments, ISIC2018 images are resized to 256×256 pixels and normalised the intensities to $[0,1]$ range.

BraTs, Brain Tumour Dataset

The Brain Tumour Segmentation Challenge provides a comprehensive and publicly available database for the development and comparison of brain tumour segmentation methods. Brain tumour segmentation is considered one of the most difficult segmentation problems in the medical domain [Isensee et al., 2021]. As the largest fully annotated database in this field, BraTS has become the standard competition for objectively evaluating segmentation techniques [Isensee et al., 2021].

The 2018 BraTS data, updated since BraTS'16, includes more clinically-acquired 3T multimodal MRI scans. The data encompasses a range of pre-operative MRI scans of Glioblastoma (GBM/HGG) and Lower Grade Glioma (LGG), all with pathologically confirmed diagnoses and available Overall Survival (OS) data, used for training, validation, and testing in the challenge. The imaging data, provided in NIFTI files, include 265 subjects, each with a range of scans (T1, post-contrast T1-weighted, T2-weighted, FLAIR) from multiple institutions all in an isotropic voxel size of $240 \times 240 \times 155$ (1 mm^3). These datasets have been manually segmented by one to four raters, following a consistent protocol, and all ground truth labels manually revised and approved by experienced neuro-radiologists [Menze et al., 2014, Bakas et al., 2017b, 2018, Bakas et al., 2017a]. In our experiments, the BraTS 2018 dataset is divided into 100 training, 25 validation, and 160 testing subjects. In this process, the three sub-compartment tumour labels are combined to segment the whole tumour, and z-score intensity normalisation ($\mu = 0$ and $\sigma = 1$) is applied to each subject and image individually.

BUS17, Breast Lesion Dataset

The BUS17 dataset, collected in 2012 from the UDIAT Diagnostic Centre of the Parc Taulí Corporation in Sabadell, Spain, used a Siemens ACUSON Sequoia C512 system with a 17L5 HD linear array transducer (8.5 MHz). It comprises 163 images from different women, averaging 760×570 pixels in size. These images feature various lesions, with 53 depicting cancerous masses and 110 showing benign lesions. Among the malignant images, there are 40 Invasive Ductal Carcinomas (IDC), 4 DCIS, 2 ILC, and 7 other unspecified malignant lesions. The benign images include 65 Unspecified Cysts (CYST), 39 Fibroadenomas (FA), and 6 other types of benign lesions. Experienced radiologists annotated the lesions in these datasets [Yap et al., 2017].

Breast cancer, characterised by uncontrolled cell division in breast tissue, often results in masses or lumps. It usually originates in the lobules or ducts of the breast. Ultrasound imaging is a crucial method for the detection of breast cancer, which uses high-frequency sound waves to create detailed images of the breast, especially useful in women with dense breast tissue. This method helps examine abnormalities found in mammography. Ultrasound is particularly effective in detecting and characterising breast mass nodes, favoured for its ease of use, non-invasiveness, low cost, and reliability in identifying mass morphology [Houssein et al., 2021]. Accurate segmentation of breast masses is vital, as the classification of a mass as malignant relies heavily on its morphological features, such as shape and boundary smoothness. Therefore, any automated system to classify breast masses must first accurately identify the mass boundary [Kumar et al., 2018].

The BUS17 dataset provides a representative sample of ultrasound images,

which are often challenging due to inherent acoustic interferences (speckle noise) and varying lesion appearances. These ultrasound images typically suffer from poor contrast and view-dependent artifacts, contributing to an intrinsic uncertainty in the analysis and segmentation of the images [Saini et al., 2010].

The dataset is divided based on the frequency of the diagnosis type, the most frequent diagnoses (IDC, FA, CYST) were selected as training and validation sets and the least frequent diagnoses (UNK, DCIS, LN, PAP, ILC, LP) were selected as testing set. The reader is referred to Figure 3.7 that illustrates the histogram plot of the BUS17 dataset categorised by lesion types. The dataset is divided into 137 training subjects, 7 validation subjects, and 19 test subjects. In this process, the images were resized to 256×192 and normalised to range $[0,1]$.

3.5.2 Results

The parameter settings and the corresponding loss function for each method and each of the datasets are listed below in Tables 3.1, 3.2 and 3.3.

The baseline architecture is Unet with 5 layers in both the encoder and decoder paths. The models are trained for 50 epochs with Adam optimiser and learning rate of 0.0001. To calculate the uncertainty map during the training and at inference time, the dropout rate and the number of stochastic forward passes were set to $d = 0.05$ and $M = 20$. The parameter settings and corresponding loss function for each method for ISIC2018, BraTS and BUS17 datasets are listed in Table 3.1, Table 3.2 and Table 3.3, respectively. All the parameters are chosen experimentally and fixed throughout the experiments.

Model	LossFunction	Parameter settings
Unet (baseline)	$\mathcal{L}_{ce}$	-
Unet+UCF	$\mathcal{L}_{UCF}$	$\alpha = 0.1, \beta = 3$
Unet+ELF	$\mathcal{L}_{ELF}$	$\alpha = 0.1, \beta = 3$
Unet+AL	$\mathcal{L}_{AL}$	$sampling = 10$
Unet+UCF+AL	$\mathcal{L}_{UCF+AL}$	$\alpha = 0.1, \beta = 3, \lambda = 1, sampling = 10$
Unet+ELF+AL	$\mathcal{L}_{ELF+AL}$	$\alpha = 0.1, \beta = 3, \lambda = 0.1, sampling = 10$

Table 3.1: Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For ISIC2018 Dataset

Model	LossFunction	Parameter settings
Unet (baseline)	$\mathcal{L}_{ce}$	-
Unet+UCF	$\mathcal{L}_{UCF}$	$\alpha = 0.1, \beta = 3$
Unet+ELF	$\mathcal{L}_{ELF}$	$\alpha = 0.1, \beta = 3$
Unet+AL	$\mathcal{L}_{AL}$	$sampling = 10$
Unet+UCF+AL	$\mathcal{L}_{UCF+AL}$	$\alpha = 0.1, \beta = 3, \lambda = 1, sampling = 10$
Unet+ELF+AL	$\mathcal{L}_{ELF+AL}$	$\alpha = 0.1, \beta = 3, \lambda = 0.1, sampling = 10$

Table 3.2: Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For BraTS Dataset

Model	LossFunction	Parameter settings
Unet (baseline)	$\mathcal{L}_{ce}$	-
Unet+UCF	$\mathcal{L}_{UCF}$	$\alpha = 0.2, \beta = 2$
Unet+ELF	$\mathcal{L}_{ELF}$	$\alpha = 0.1, \beta = 3$
Unet+AL	$\mathcal{L}_{AL}$	$sampling = 10$
Unet+UCF+AL	$\mathcal{L}_{UCF+AL}$	$\alpha = 0.2, \beta = 2, \lambda = 1, sampling = 10$
Unet+ELF+AL	$\mathcal{L}_{ELF+AL}$	$\alpha = 0.1, \beta = 3, \lambda = 0.1, sampling = 10$

Table 3.3: Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For BUS17 Dataset

Table 3.4 shows the quantitative measurements of six different models: Unet (baseline), AL, UCF Loss, ELF Loss, UCF+AL and ELF+AL (described in Appendix A) using test subjects from ISIC2018, BraTS and BUS17 datasets. To evaluate the statistical significance of the values for each performance measure, we applied the non-parametric Wilcoxon signed rank statistical test [Gibbons and Chakraborti, 2011]. Tables A.1, A.2 and A.3 in Appendix A show the p-values for the corresponding pairwise tests.

The results based on Table 3.4 show that the model with ELF and the model with ELF+AL have the highest dice compared to other models for the ISIC2018 dataset. Similarly for the BraTS dataset, the model with ELF loss achieved the highest dice compared to other models. The Unet+ELF loss results are statistically significant, except when compared with Unet+UCF loss. For the BUS dataset, Unet+ELF+AL has the highest dice, however, the results are relatively close and the statistical test did not show any significant difference between the compared models.

The state-of-the-art segmentation methods for the BraTS 2018 challenge have achieved a Dice coefficient of 0.90 [Ghaffari et al., 2019]. For the ISIC2018 dataset, the best Dice score reported in the literature is 0.902 [Yu et al., 2023], and for the BUS17 dataset, the highest reported Dice score is 0.842 [Yeung et al., 2023].

However, it is important to note that the training, validation, and testing sets chosen for these experiments in this thesis are different from those used in the cited literature. This difference in datasets means that direct comparisons with the state-of-the-art results are not entirely fair or accurate. Additionally, the primary goal of our experiments is not to surpass the best Dice scores reported in the state-of-the-art methods. Instead, we used a basic architecture without incorporating advanced techniques such as data

augmentation, ensemble methods, or post-processing improvements. Our focus is primarily on evaluating how the uncertainty-aware loss design impacts the interpretation of segmentation result uncertainty.

The purpose of these experiments is to investigate whether the proposed uncertainty-aware loss design can enhance the understanding of uncertainty in segmentation results, which is a critical aspect of medical image analysis. While achieving competitive Dice scores is desirable, our main contribution lies in providing a framework that can improve the reliability and interpretability of segmentation outputs by addressing uncertainty.

Dataset	Model	Dice
ISIC2018	Unet+Cross Entropy Loss (Baseline)	0.8872± 0.014
	Unet+Aleatoric Loss (AL)	0.8896± 0.015
	Unet+UCF Loss	0.8863± 0.015
	Unet+ELF Loss	0.8917± 0.012
	Unet+UCF + AL	0.8891± 0.016
	Unet+ELF + AL	0.8917± 0.013
BraTS	Unet+Cross Entropy Loss (Baseline)	0.8634± 0.019
	Unet+Aleatoric Loss (AL)	0.8671± 0.014
	Unet+UCF Loss	0.8693± 0.013
	Unet+ELF Loss	0.8710± 0.014
	Unet+UCF + AL	0.8667± 0.014
	Unet+ELF + AL	0.8688± 0.014
BUS	Unet+Cross Entropy Loss (Baseline)	0.5430±0.167
	Unet+Aleatoric Loss (AL)	0.5438±0.148
	Unet+UCF Loss	0.5501±0.179
	Unet+ELF Loss	0.5511±0.155
	Unet+UCF + AL	0.5669±0.161
	Unet+ELF + AL	0.5715±0.157

Table 3.4: Segmentation performance comparison in terms of dice coefficients for different models on ISIC2018, BraTS whole tumour and BUS17 datasets on a test subjects for baseline Unet, Unet with AL, Unet with UCF loss, Unet with ELF loss, Unet with the UCF loss and AL as regularisation term and Unet with ELF loss and AL as regularisation term.

The performance variation of proposed loss functions such as the UCF loss

across different datasets, shows the complex and data-specific nature of medical image segmentation. This variation highlights the importance of tailoring loss functions and model architectures to the unique characteristics of each medical imaging task. Medical images differ significantly in terms of resolution, contrast, and the nature of the anatomical structures being segmented.

To further investigate the effects of the proposed methods, we performed visual inspections of the prediction results of test subjects. In the comparative analysis of various models on the whole tumour segmentation of BraTS dataset, several key observations emerge. The baseline model, model with AL, model with UCF, model with ELF and their combinations with AL as a regularisation term were examined. In Figure 3.11, the model with ELF loss, although it achieved the highest average dice overall, in this example, it struggled in certain regions (shown by letter (*T*) in Figure 3.11). These areas were not segmented effectively and the corresponding model uncertainty heatmaps for the model with ELF loss exhibited high confidence, indicating a possible oversight in recognising these regions. In contrast, the UCF model demonstrated a more adept response in segmenting these challenging regions, in addition to effectively reflecting the epistemic uncertainty in the heatmap. The combination of UCF with aleatoric loss did not improve the identification of regions and the performance of uncertainty heatmaps. The analysis underscores the varying effectiveness of these models and combinations in handling uncertain and challenging segmentation tasks within the BraTS dataset.

In this visual inspection of model performance on the BraTS dataset shown in Figure 3.12, as one of the complex cases, several critical observations were noted. The primary objective for the models in BraTS dataset is tumour segmentation. However, the challenge of segmentation intensifies when

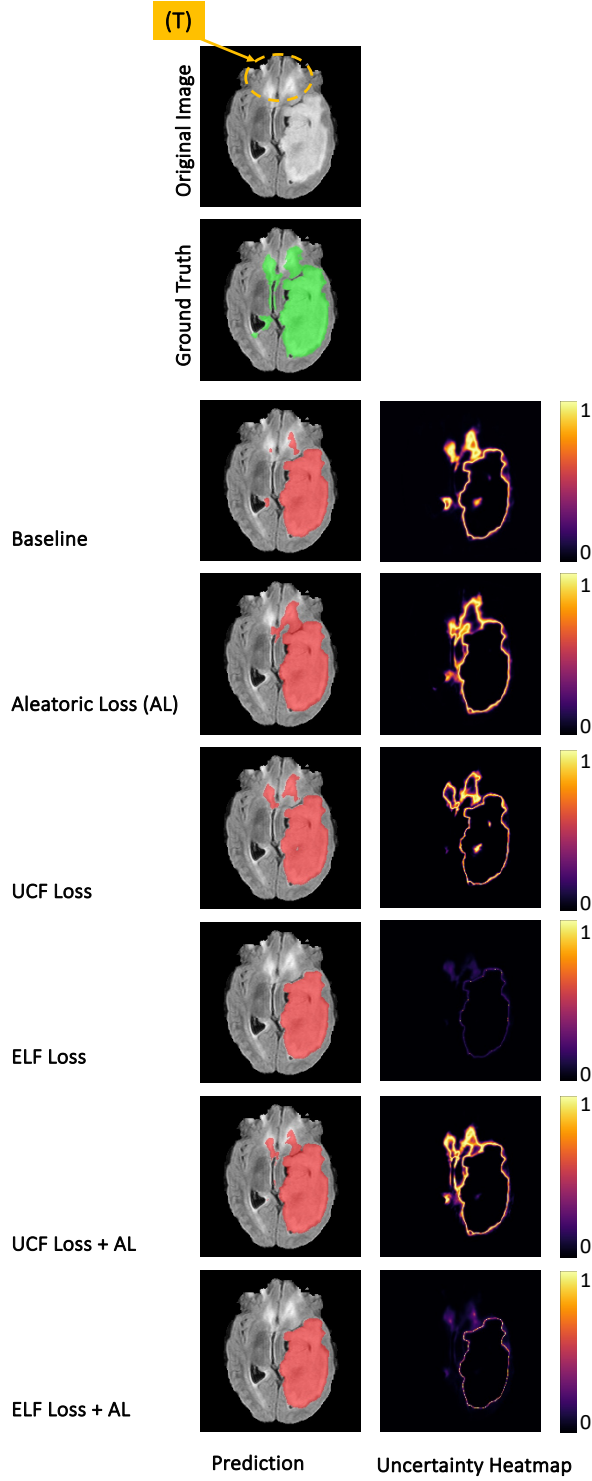


Figure 3.11: Comparative analysis of various models on the BraTS dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value. Letter T represents only a specific region of the brain and part of the whole tumour.

subjects have additional brain abnormalities that are not annotated as tumours, as these unknown elements can confuse the segmentation model. One such case involved a subject exhibiting a tumour on the left side of the image, which aligns with the annotated region of interest as shown in Figure 3.12 by letter (*T*). Interestingly, the subject’s T2 FLAIR contrast imaging also revealed hyperintensity regions on the right side of the image and abnormal white matter tissue, as shown in Figure 3.12 by letter (*A*)¹. These findings, while potentially clinically significant, fall outside the model’s designated task of tumour segmentation. In such scenarios, there are two preferable outcomes for the model’s behaviour: either segmenting these non-tumour regions with high uncertainty, suggesting a need for further investigation, or correctly not segmenting these areas as tumour tissue with high confidence.

The performance of different models varied in this context. The ELF model successfully disregarded the seemingly abnormal white matter region but mistakenly identified the hyperintense area on the right side of the image as tumour tissue with high confidence. The addition of AL to the ELF model (ELF+Aleatoric) showed slight improvement by more distinctly highlighting the boundaries of these regions. Meanwhile, both the Aleatoric and UCF loss models incorrectly segmented these non-tumour regions as tumour, though they displayed a significantly higher degree of uncertainty in these areas compared to the actual tumour on the left side. The combination of UCF and AL offered a marginally better interpretation, balancing segmentation accuracy with more interpretable uncertainty heatmap. This analysis underscores the complexities and nuances in model performance,

¹As highlighted in section 3.5.1, it is crucial to acknowledge that the BraTS datasets were manually segmented by one to four raters who adhered to a uniform protocol. Furthermore, all ground truth labels were meticulously reviewed and confirmed by expert neuroradiologists, thus addressing obvious concerns regarding the precision of the annotations.

particularly in cases with additional brain abnormalities that challenge the segmentation task. The model's ability to differentiate between tumour and non-tumour regions, and to express uncertainty in ambiguous cases, is crucial for reliable and clinically relevant image analysis.

Turning to the skin lesion data, this particular evaluation shown in Figure 3.13 is a specific subject for which, dermoscopic image presented unique challenges, including the presence of colour charts and hair, which are common sources of aleatoric uncertainty in such datasets. The model with ELF loss successfully segmented the skin lesion without being confused by the colour charts, demonstrating high confidence in its analysis. Similarly, when combined with AL (ELF+AL), the model maintained this accuracy. However, the addition of AL interestingly intensified the uncertainty of the model around the lesion's boundaries. The model with UCF loss, when operating independently, showed a vulnerability to the sources of aleatoric uncertainty. It incorrectly segmented some non-lesion areas, mistaking them for part of the lesion, although it correctly marked these regions with high uncertainty values in the heatmap. This outcome underscores the model's ability to recognise its limitations in segmenting under complex conditions. However, the combination of UCF with AL became a particularly effective approach. This hybrid model not only resolved the wrong segmentations triggered by aleatoric uncertainty but also demonstrated robustness against challenges like similar intensity in the image.

Figure 3.14 illustrates a comparison of six various tailored loss functions using BUS17 dataset, specifically on a subject with a UNK lesion type classified as malignant. As shown in Figure 3.14 all models correctly identified the UNK lesion. However, the model with just aleatoric loss mistakenly segmented other similar intensity regions as lesions, often with higher uncertainty than the actual lesion. In contrast, the UCF and ELF models

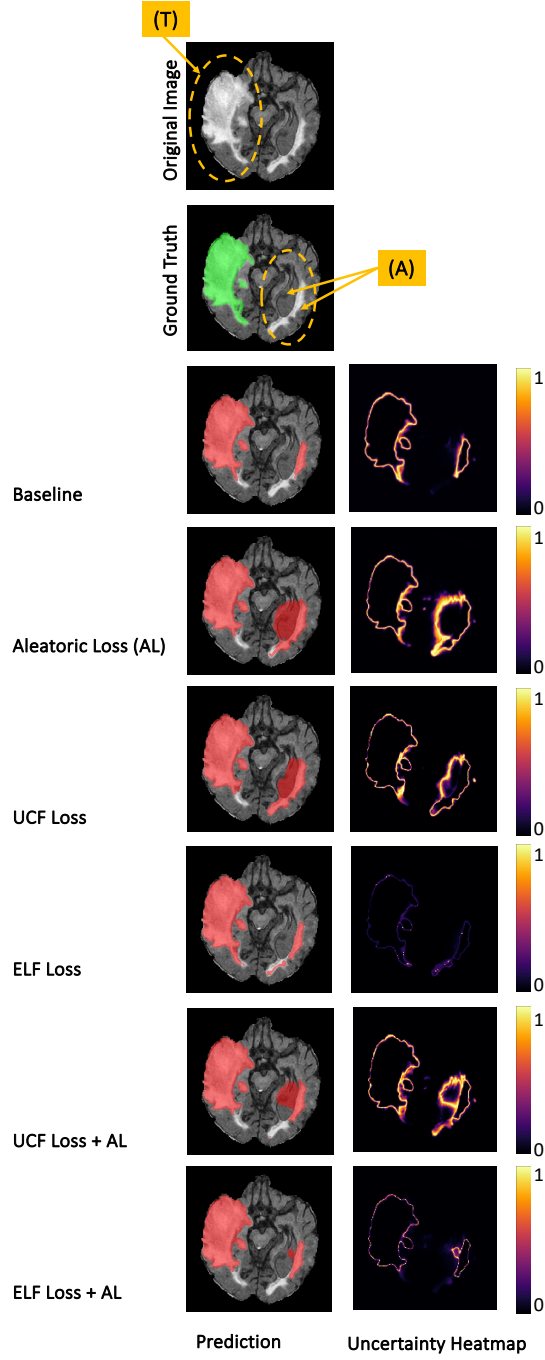


Figure 3.12: Comparative analysis of various models on the BraTS dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value. Letter T represents only a specific region of the brain with the whole tumour. Letter A represents only a specific region of the brain with the apparent abnormality that was not annotated as tumour.

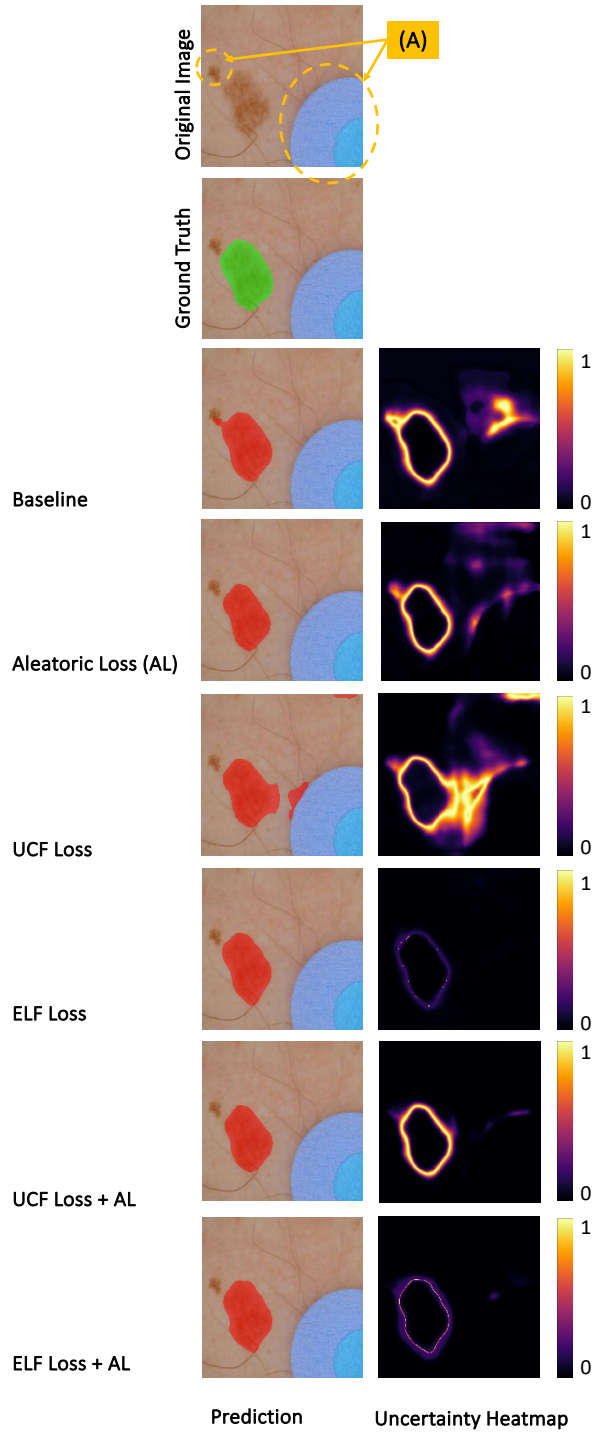


Figure 3.13: Comparative analysis of various models on the ISIC2018 dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value. Letter *A* represents specific regions with artefacts.

achieved more accurate segmentation and provided a more coherent interpretation of uncertainty, generally attributing lower uncertainty to the background in comparison to the aleatoric loss model. The ELF model, when combined with aleatoric loss, generated a more informative uncertainty heatmap (showing higher uncertainty in the boundary of the lesion specifically on the right side of the lesion), while still maintaining a high level of segmentation accuracy.

The rationale behind the BUS17 data split, using the most frequent lesion types for training and the least common for testing, was to challenge the models with previously unseen lesion types, effectively treating these test cases as OOD data. This was expected to introduce high epistemic uncertainty, providing an opportunity to assess the performance of each uniquely tailored loss function and its consequent impact on the interpretation of uncertainty heatmaps. The resulting analysis has proven insightful, highlighting how models adapt to unfamiliar data and the crucial role that tailored uncertainty management plays in model reliability, particularly in a medical setting where accurate diagnosis is important.

It is important to note that the effectiveness of UCF, ELF, and AL regularisation term could vary depending on the characteristics of the data and the dataset used for training. For datasets with high noise or variability, AL is important as it helps the model become more robust against inherent data imperfections. UCF, on the other hand, is more beneficial when dealing with high epistemic uncertainty, where the model lacks knowledge or exposure to varied examples. In contrast, ELF focuses on iterative error correction and can enhance performance by learning from previous mistakes. The interplay between these loss functions is tailored to the specific data context and in some scenarios the effect of AL could cause too much regularisation the results in underfitting.

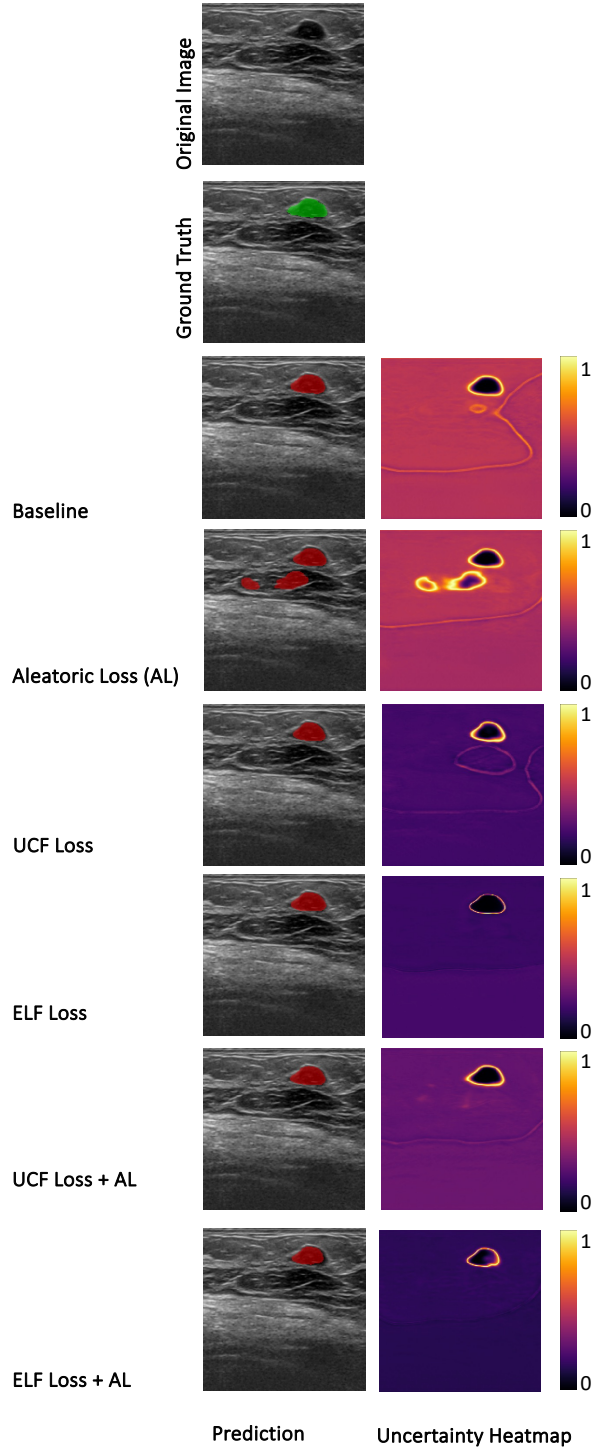


Figure 3.14: Comparative analysis of various models on the BUS17 dataset, the baseline model, model with AL, model with UCF, model with ELF, and their combinations with AL as regularisation term, UCF+AL and ELF+AL. The black colour in the heatmap indicates the lowest uncertainty value and the yellow colour shows the highest uncertainty value.

3.6 Discussion

Our research demonstrates how different factors across medical images of different modalities influence epistemic and aleatoric uncertainty in deep medical image segmentation. These findings underscore the complex nature of uncertainties in clinical scenarios, and the significant role of uncertainty maps generated by DCNNs. These maps are likely to be important in enhancing diagnostic accuracy and patient care, as they could provide valuable insight into the confidence level of the model’s predictions.

Complementing this, recent studies have highlighted that while deep learning systems have improved in accuracy, their robustness, especially in clinical settings, remains a concern. Uncertainty measures have emerged as a promising direction to improve failure detection in these systems. However, evaluations of common voxel-wise uncertainty measures reveal that they tend to be well-calibrated at the dataset level but miscalibrated at the subject level, thereby compromising the reliability of uncertainty estimates [Jungo and Reyes, 2019]. This finding is particularly relevant as it highlights the importance of developing subject-wise uncertainty estimations, ensuring that the uncertainty measures are not just globally accurate but also reliably reflect the uncertainties at an individual patient level.

This further underscores the importance of our findings, particularly from a technical perspective. By exploring how various image factors influence epistemic and aleatoric uncertainties, our study sheds light on the critical role of uncertainty-aware models in medical imaging. These models are useful in enhancing the interpretability and reliability of uncertainty maps generated by DCNNs. We demonstrated in our research how various tailored loss functions handle uncertainty differently and that the design of a correctly-tailored loss is pivotal for ensuring consistency and accuracy in

the interpretations derived from the uncertainty maps.

By offering a detailed examination of the factors contributing to uncertainty in medical image segmentation, our work provides a clear path for developing more robust and reliable medical imaging techniques. The careful design of the model loss function tailored for the task at hand and careful validation of the performance and interpretability are crucial for the future of medical diagnostics, where the precision and consistency of image interpretation can have profound implications for patient care and treatment outcomes. The move towards subject-level uncertainty estimation is a key aspect of this progress.

Holistic Approach to Interpretation Integrating patient-specific factors, such as disease biomarkers, age, gender, and other clinical data, is also likely to be essential in interpreting uncertainty maps. This holistic approach ensures that interpretations are not solely reliant on the model output, but are enriched by a comprehensive understanding of the patient’s health profile. For example, high uncertainty in tumour segmentation might be more comprehensively understood considering the patient’s specific disease biomarkers or age. This integration is crucial for accurate and personalised medical decisions.

The context provided by additional patient information is invaluable in interpreting the uncertainty of segmentation results [Simpkin and Armstrong, 2019]. For example, a DCNN model showing high uncertainty in segmenting a tumour could be indicative of the presence of an atypical disease, especially when correlated with other patient information. In contrast, low uncertainty in a model prediction might be re-evaluated if it contradicts the patient’s clinical symptoms or disease biomarker profile.

Model Limitations and Future Directions Recognising the limitations of current DCNN models is crucial. This awareness can guide clinicians in critically interpreting uncertainty, especially in cases that deviate from typical training scenarios. Future research should focus on developing models that incorporate a wider spectrum of patient data types, thus improving accuracy and reducing uncertainty in segmentation tasks, while also making the interpretation of this uncertainty clearer and more straightforward. Such advances are key to the evolution of personalised medicine, where the combination of advanced imaging techniques and comprehensive patient data can lead to more tailored diagnostics and treatment plans.

3.7 Conclusion

Medical image segmentation is a challenging task due to the complexity of the data and the complex nature of uncertainty in the images. In this Chapter, we focused on the interpretability of uncertainty estimation in Bayesian Deep Convolutional Neural Networks, specifically within the context of medical image segmentation. We have chosen datasets that contain a wide variety of lesion shapes, border irregularities, inconsistencies in acquisition methods, variance in lesion texture, and the presence of image artefacts. We introduced tailored loss functions including an epistemic uncertainty-based attention map on cross-entropy loss (UCF), an error feedback weight map on cross-entropy loss (ELF), and their combinations with aleatoric loss as regularisation term, UCF+AL loss and ELF+AL loss. We used each of these loss functions for training a fully supervised DCNN model. Our focus was to propose and compare the tailored loss functions that handle uncertainties of the model and data uncertainties in different ways.

We designed comprehensive experiments on three benchmark medical imaging datasets across different modalities, each having their unique challenges and characterisation for analysing and evaluating the proposed loss functions and their effect on the interpretation of uncertainty heatmaps.

Among the analysis of segmentation models applied to the ISIC2018, BraTS, and BUS17 datasets, the ELF loss, particularly when combined with AL (ELF+AL), stands out by achieving the highest Dice scores, indicating better segmentation for the ISIC2018 dataset. This trend holds for the BraTS dataset, with the ELF model outperforming other loss functions. Visual inspections further validate these quantitative findings, revealing that while models like UCF show proficiency in interpreting epistemic uncertainty, the addition of AL refines the models' precision, as evidenced by the more discerning uncertainty heatmaps of ELF+AL.

It is important to note that the observed disparity between the Dice score and visual inspection in our experiments for ELF and UCF loss functions highlights an important aspect of medical image segmentation which is the need to balance quantitative metrics with qualitative assessments of model performance. While the ELF model achieves higher Dice scores, indicating better segmentation accuracy on a pixel-level basis, the UCF model, despite its lower Dice scores, demonstrates more reliable performance upon visual inspection. This suggests that while ELF is effective at optimising the Dice metric, it may not fully capture the nuances of image segmentation that are vital for clinical interpretation, such as accurately delineating complex lesion boundaries.

This finding underscores the importance of not solely relying on standard metrics like the Dice score but also considering how well a model's outputs align with clinical expectations and how uncertainty is communicated. In

medical imaging, where decisions have major effects on patient outcomes, the reliability of the model, as evidenced through visual inspection and uncertainty mapping, are as critical as quantitative accuracy. This highlights the necessity of developing and validating tailored loss functions that prioritise not just segmentation accuracy but also the practical utility and interpretability of the resulting uncertainty maps in real-world clinical settings.

This blend of tailored loss functions, especially in handling out-of-distribution data in the BUS dataset, underscores the efficacy of integrating tailored uncertainty management strategies. Such integration is pivotal in medical diagnostics, where the accuracy of lesion segmentation and the interpretation of uncertainty directly affect clinical decisions.

We demonstrated in this chapter, that various tailored loss functions handle uncertainty differently and the design of a tailored loss is pivotal for ensuring consistency and accuracy in the interpretations derived from the uncertainty maps. The careful design of the loss function tailored for the task at hand and careful validation of the performance and interpretability are crucial for the future of medical diagnostics, where the precision and consistency of image interpretation can have profound implications for patient care and treatment outcomes.

Chapter 4

Uncertainty-Guided and Spatially-Constrained Weakly Supervised Deep Medical Image Segmentation

4.1 Context

In the preceding chapters, we investigated the intricacies of medical image interpretation and the various challenges it poses, particularly in the context of MRI, dermoscopy, and other complex imaging modalities. In Chapter 2, we highlighted the issues of distribution drift, the scarcity of centralised open-source databases, and the long-tailed distribution of diseases, which complicates the task of image annotation and segmentation. In chapter 3, we explored how manual segmentation, while precise, is labour-intensive and susceptible to biases, and how automatic segmentation techniques, though promising, still grapple with the unique challenges posed

by the diverse and intricate nature of medical imaging data.

Furthermore, we discussed the complexities involved in handling noisy and sparse labels in deep learning, underscoring the need for robust strategies to manage uncertainties and inaccuracies in segmentation masks. As we move into this chapter, we will be looking at how to tackle these issues by using *weakly supervised learning*, particularly with bounding box annotations.

In this chapter, we propose a weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty-guided and spatially-constrained loss, which only requires simple and fast bounding box annotations for model training. The contributions in this chapter are as follows:

- We propose predictive uncertainty estimation during training to guide the model learning from the image region with high predictive confidence.
- Additionally, a CRF-based local spatial constraint is incorporated to the loss function, which regularises the predicted labels of a local region. This CRF loss term is independent of the training labels (bounding box annotations here), which prevents the model overfitting to the bounding box annotation.
- We evaluate our method on a public dermoscopic dataset (ISIC2018) containing different types of skin lesions. Our method outperforms state-of-the-art learning-based (DeepCut) and non-learning based (GrabCut) methods, using dice coefficient as an accuracy metric.

The results of this chapter has been published at the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) [Mahani et al., 2022].

4.2 Introduction

The challenge of having access to large datasets with accurate annotations is significant in medical imaging [Karimi et al., 2020]. In many real world clinical applications, inaccurate annotations are obtained for training machine learning models due to the following reasons: (a) for some clinical problems, object segmentation is not trivial even for a trained human expert [Ibtehaz and Rahman, 2020]; (b) accurate labelling of medical images is very resource-intensive as it requires a domain experts’ knowledge [Karimi et al., 2020]; (c) the accurate annotation can be particularly hard for very small objects; (d) detailed annotation is a time-consuming process (e) it is subject to human errors and fallibility; and (f) in some medical applications, significant inter-observer variability necessitates obtaining consensus labels or annotating from several experts aggregating the annotations using appropriate methods [Karimi et al., 2020].

Although in some studies, a large number of medical image datasets were annotated by experts, such studies require a large amount of financial and logistics not available for many of the medical domains. On the other hand, for successful deep learning methods in medical applications having a large dataset with reliable annotations is crucial [Karimi et al., 2020].

Over recent years, a number of weakly-supervised CNNs have emerged for semantic segmentation of natural images. However, a small number of these methods have been specifically tailored for segmentation tasks in medical imaging [Yang et al., 2020]. These successful applications have shown that weakly supervised learning methods are able to reduce the need for pixel-level annotations. Various weak supervision medical image segmentation methods have been performed using different forms of annotations, such as points [Kervadec et al., 2019, Qu et al., 2019], image-level annotations [Wu

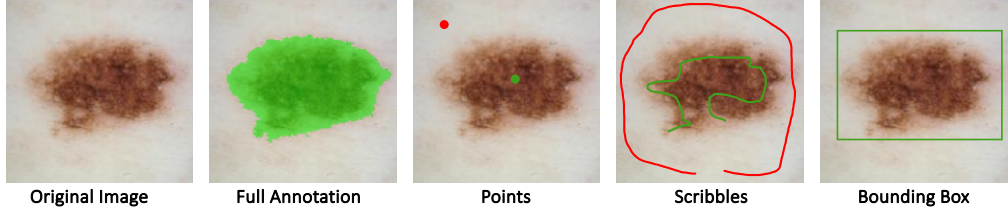


Figure 4.1: The differences between full annotation, points, scribbles, and bounding box annotation on a skin lesion dataset sample. Data sample is selected from ISIC2018 Skin Lesion Dataset.

et al., 2019] and scribbles [Can et al., 2018, Liu et al., 2022] and bounding boxes [Rajchl et al., 2016, Yang et al., 2018, Kervadec et al., 2020]. Figure 4.1 shows the differences between full annotation, scribbles, points, and bounding box annotation on a skin lesion dataset sample. Note that the ‘image-level annotation’ usually refers to binary labels indicating if lesion tissues exist in the image.

Weakly supervised segmentation methods using *image-level annotations* achieve fine-grained segmentation with minimal expert input [Wu et al., 2019]. These approaches, primarily leverage Class Activation Maps (CAMs) [Zhou et al., 2016] and Multiple Instance Learning (MIL) [Jia et al., 2017]. CAMs, a critical component in many such methods, utilise a classification network’s intermediate outputs to generate segmentation masks, essentially pinpointing the regions of interest in an image that contribute most significantly to the classification decision. On the other hand, MIL approaches the problem of weak supervision as one of label noise, focusing on the classification of sets of instances rather than individual ones. Such methods have demonstrated their potential in enhancing model performance compared to standard benchmarks in weakly supervised segmentation using image-level annotation [Tajbakhsh et al., 2020]. While image-level annotations provide a broad overview for segmentation, the next approach, point annotations, introduces a more focused and localised method of guiding segmentation.

Weakly supervised segmentation methods using *point annotations* were proposed for balancing the need for precision in object class identification with the practical limitations of exhaustive pixel-wise labeling [Bearman et al., 2016]. The process works by assigning known classes to a select set of points within an image, effectively using these points as guides for the segmentation of the entire image.

The first method proposed for using point annotations in weakly supervised segmentation [Bearman et al., 2016] employed a modified loss function that accounts for the relative importance of each annotated point. This importance is variable and can be determined through several strategies: it might be uniform across all points, based on the confidence of multiple annotators, or weighted according to the sequence in which points are presented, with early points typically representing more prominent object features. In practice, the point annotation itself is not sufficient to directly supervise the training of neural networks. Studies that use point annotations typically incorporate some kind of prior [Chen et al., 2020b], for example, using the original image and a shape prior [Qu et al., 2019] or introducing an ‘objectness’ prior (a pre-trained measure that estimates the likelihood of each pixel belonging to any object class versus the background [Bearman et al., 2016]). In essence, point annotation-based weakly supervised segmentation refines the training of CNNs for semantic segmentation by efficiently utilising sparse but strategically important point annotations and leveraging additional context through priors.

Scribble annotation emerged as a natural progression from point-based annotation in weakly supervised segmentation, primarily in response to the limitations and challenges associated with point annotations. While point annotation offers a simpler alternative to full region masks and is easier to generate, it demands annotations for the entire dataset and is highly

sensitive to the exact location of the points [Lee and Jeong, 2020]. This requirement can be labour intensive and less practical for large datasets, or for certain types of datasets such as the histology dataset, for which the task is to segment a significant number of nuclei in each image [Lee and Jeong, 2020].

Scribble labels, which are loosely drawn lines or marks over the objects of interest, offer a faster and less precise method compared to detailed point annotations, as shown in Figure 4.1. These scribbles provide enough guidance for regularised networks and standard segmentation techniques to produce a detailed segmentation. Some methods also introduce additional model parameters or techniques such as boundary prediction and adversarial training to refine the outcomes further [Lee and Jeong, 2020]. Others combined pseudo-labeling and label filtering to generate reliable labels under weak supervision. The pseudo labels were improved by leveraging the consistency of the predictions during iterations [Liu et al., 2022].

Although using scribble annotations for training can save time for expert segmentation, it may compromise the accuracy of segmentation models due to their imprecise nature and insufficient supervised signals, specifically with the quality defects common in medical images [Li et al., 2023]. Scribble annotations, with their relaxed approach compared to point annotations, pave the way for exploring even broader and more user-friendly annotation techniques, such as bounding box annotations, which offer a different balance of precision and ease.

Among the different forms of weak annotations, *bounding boxes* have become very popular. Bounding box annotations have been widely used in interactive segmentation approaches such as the GrabCut method [Rother et al., 2004]. Furthermore, they are simple, low cost as they require light

storage and allow the problem to be spatially constrained [Kervadec et al., 2019, 2020]. Figure 4.1 shows an example of bounding box annotation for skin lesion segmentation task.

A common strategy in learning from weak annotations is the *proposal-based* method. This approach generally involves using initial, rough annotations, often referred to as pseudo-annotations or proposals, and then iteratively refining and updating these annotations as the training progresses. As shown in Figure 4.2a, this method starts with a basic approximation of the target region—in this case a rectangular bounding box for the skin lesion segmentation task—and progressively adjusts this initial guess to more accurately represent the true segmentation targets through a series of updates. This iterative refinement aims to enhance the accuracy and reliability of the model’s learning process from weakly labeled data. However, one of the drawbacks of such strategies is that proposals contain errors, which might be propagated through the training. This could affect segmentation performance and lower the chance of convergence. The error propagation could be more noticeable when working on medical images as the targets are normally challenging and have poor contrasts. Moreover, the iterative proposal generation increases the computational costs of training Kervadec et al. [2019, 2020]. Our initial experiments for lesion segmentation using bounding box annotations support the error-prone nature of proposal-generation-based methods. We found that approaches that require proposal generation and refinement, are typically hard to converge and it is very hard to tune the parameters to generate proposals that contain fewer errors.

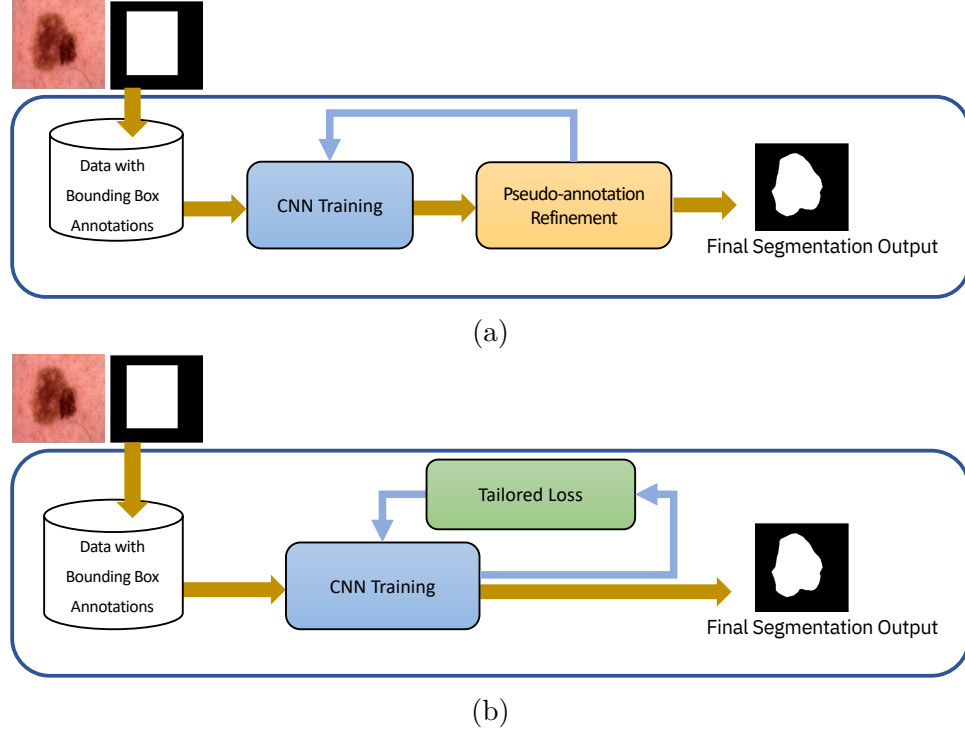


Figure 4.2: (a) Proposal-based weakly supervised medical image segmentation strategy, that uses weak annotations to generate pseudo-annotations known as *proposals*. These labels are generated iteratively to refine the parameters of deep CNNs to mimic the full supervision. (b) Weakly supervised medical image segmentation strategy using a tailored loss function that does not require iterative refinement of pseudo annotations.

As mentioned earlier, despite the increasing number of studies focusing on weakly-supervised DCNNs in semantic segmentation [Kervadec et al., 2019], due to the difficulty of leveraging weak annotations in medical imaging, the number of compelling studies that use bounding box annotations, in particular, is limited. In this chapter, we briefly discuss some of the relevant studies.

The GrabCut method is an interactive foreground/background segmentation in still images, crucial for practical image editing tasks, standing out as a non-learning based approach. By iteratively updating the parameters of a Gaussian Mixture Model (GMM) [Reynolds et al., 2009] and optimising the resulting energy with a graph cut [Boykov and Jolly, 2001], GrabCut

combines both texture (color) and edge (contrast) information to achieve precise segmentation. This iterative optimisation allows for a more powerful and nuanced approach to image segmentation, significantly simplifying the user interaction needed for high-quality results [Rother et al., 2004].

In [Kervadec et al., 2020] a novel, weakly-supervised learning approach for segmentation is introduced that utilises global constraints derived from box annotations, specifically applying a classical tightness prior by imposing constraints on network outputs, ensuring that each line within the bounding box contains at least one foreground pixel. This method, demonstrates superior performance over well established methods such as DeepCut, eliminating the need for computationally intensive proposal generation.

Another study presents a weakly-supervised CNN specifically designed for renal tumour segmentation, leveraging bounding boxes as a form of weak annotation to train the model [Yang et al., 2020]. The approach incorporates CRF to generate pseudo masks from bounding boxes, followed by a novel group training strategy where multiple CNNs are trained with these pseudo masks and then fused to refine the segmentation. Finally, a voxel-wise weighted cross-entropy loss function is utilised to further enhance the training process, ensuring precise and robust segmentation outcomes. This technical framework not only accelerates the training procedure but also effectively mitigates common issues like under-segmentation, ultimately yielding high accuracy on a diverse clinical dataset of renal tumours [Yang et al., 2020].

In [Wang and Xia, 2021], a weakly supervised image segmentation method is introduced utilising tight bounding box annotations and a novel approach of generalised MIL coupled with smooth maximum approximation for integrating bounding box tightness priors into a deep neural network. The

method defines positive and negative bags through parallel crossing lines and individual pixels outside bounding boxes, respectively, and employs α -softmax and α -quasimax functions to address numerical instability [Wang and Xia, 2021].

DeepCut [Rajchl et al., 2016] is one of the mostly cited bounding box-based weakly supervised method. It is inspired by GrabCut method and uses the common strategy of generation and refinement of pseudo-masks known as *proposals* during the training. These pseudo-masks are generated iteratively to refine the parameters of DCNNs to mimic full supervision. The pseudo-masks are updated in a post-processing step using dense Conditional Random Field [Krähenbühl and Koltun, 2011] for the next training iteration. Although the DeepCut method outperformed the conventional GrabCut, another study showed that the pseudo-masks may gradually disappear in some cases. This suggests that DeepCut could be very unstable during training [Kervadec et al., 2019]. Moreover, despite using Dense CRF to smooth the predicted pseudo-masks, it seems that the strategy of using CRF as a separate post-processing step failed to prevent the bounding box annotations from shrinking and generating empty predictions.

DeepCut is a good example of a potentially powerful, but error-prone proposal-based method. The error propagation in such methods could be more noticeable when working on medical images as the targets are normally challenging and have poor contrasts. Moreover, the iterative proposal generation increases the computational costs of training [Kervadec et al., 2019, 2020]. Considering these points, it would be preferable to use strategies that do not require proposal generation and updates.

Our aim in this chapter, as shown in Figure 4.2b, is to propose strategies that *do not* require iterative pseudo-annotation refinement. Instead, we

utilise the spatial constraints that the bounding box annotation provides. Furthermore, our hypothesis is that epistemic uncertainty estimation could help the training to focus on the regions inside the bounding box that are most certain.

4.3 Methodology

An overview of our proposed method is shown in Figure 4.3. The input to the framework is the original images and their corresponding bounding box annotations. The proposed loss function consists of a Weighted Cross Entropy (WCE) term and a CRF term. The pixel-wise weighting map of the WCE term is derived from the epistemic uncertainty estimation (explained in Chapter 3) and combined with the bounding box background constraint (any pixel/voxel outside the bounding box is considered as background). The CRF term helps to regularise the predicted labels in a local region, which is based on the original image intensity and is independent of the bounding box annotation. The loss function is optimised using a widely used encoder-decoder DCNN method. For model deployment, only the original image is needed as the input, the trained model is then able to predict a detailed object region rather than a bounding box. Each of the key components is described in the following subsections.

4.3.1 DCNN-based image segmentation

The architecture of our proposed method is implemented based on a well-known encoder-decoder network with skip connections (Unet [Ronneberger et al., 2015]). The basic architecture of our DCNN consists of a multi-layer encoding path to capture the image context and a symmetric decoding

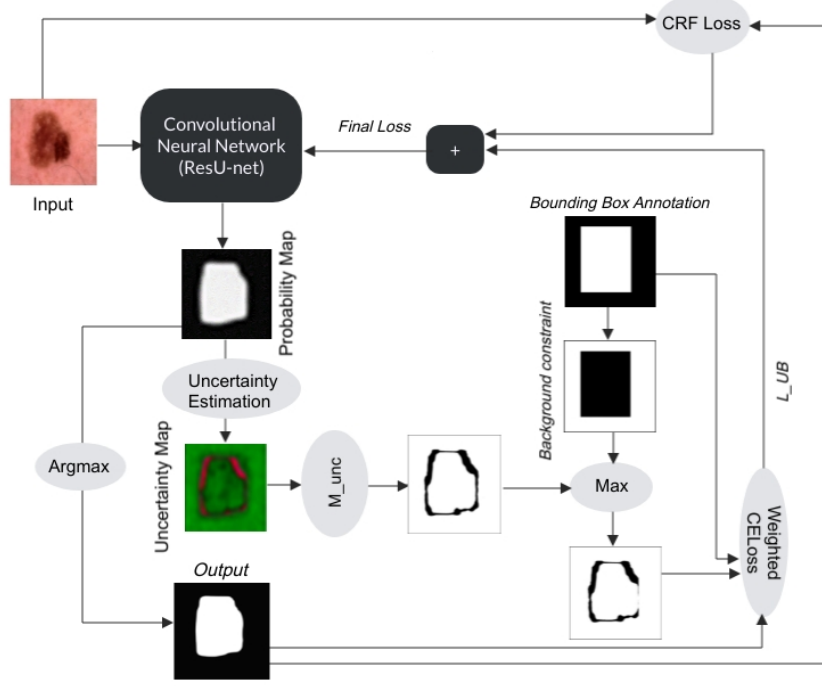


Figure 4.3: Overview of the proposed bounding box based weakly supervised framework. We incorporated predictive uncertainty estimation during the training process to help the model focus on learning from image areas where it is more confident in its predictions, while also considering a background constraint. Furthermore, we included a local spatial constraint based on CRF in the loss function to enforce consistency in the predicted labels within a specific region.

path that enables the production of high-resolution segmentation results. The residual block was proposed first in the ResUnet architecture for classification problems [He et al., 2016a]. The main idea of ResUnet was to take advantage of shortcut connections to address the problem of vanishing gradients which may hinder the convergence of the network. Using these shortcut connections does not add extra parameters or complexity to the network. Due to the success of ResUnet, several architectures are proposed as variants of this architecture [Huang et al., 2016, Xie et al., 2017, He et al., 2016b, Huang et al., 2017]. The structure of a residual block is to have a shortcut connection that skips one or more layers. The shortcut connection performs identity mapping, and is added to the outputs of the skipped layers [He et al., 2016a]. To take advantage of the benefits

of Residual networks, in our experiment we chose the residual version of Unet.

Instead of the conventional layers of Unet, the architecture is constructed using Residual blocks containing, two 3×3 convolutions each followed by a spatial dropout [Tompson et al., 2015], a batch normalisation [Ioffe and Szegedy, 2015] and ReLU, and the block contains one skip connection that links the input of the block to the output (ResUnet [Zhang et al., 2018]).

4.3.2 Uncertainty guided loss and background constraint

We propose to utilise uncertainty concepts to guide the model training process. The intuition behind the design is that the target objects in different images share similar characteristics. During the model training process, these common object regions are likely to be predicted correctly with high confidence and certainty. Hence we propose to use an uncertainty measure to locate these regions. The model is then able to focus on learning from the image regions with *high* certainty rather than from the whole bounding-box region. We follow the method proposed by Gal et al. [Gal, 2016] for uncertainty estimation. Predictive entropy, as detailed in [Gal, 2016], serves as an indicator of the model’s uncertainty, quantifying the predictive distribution’s information content. The predictive entropy $unc = H[y|x, \mathcal{D}]$ (as outlined in equation 3.1 from Section 3.3.2) reflects epistemic uncertainty, with values ranging from 0 to 1, where higher values denote greater uncertainty in the prediction. For an in-depth explanation of how epistemic uncertainty is computed, readers can refer back to Section 3.3.2.

Our proposal is to utilise the estimated uncertainty map as a weight map

on a per-pixel basis. This weight map will be used to calculate a weighted cross entropy loss, taking into account both the predicted output and the bounding-box annotation. The idea is to assign a higher weight to pixels with lower uncertainty (higher certainty), which will result in a larger contribution to the overall loss calculation. Therefore, the weight map based on uncertainty, denoted as M_{unc} , is formulated as shown in Equation (4.1),

$$M_{unc}(x) = \begin{cases} 1 & \text{if } unc < \theta \\ 0 & \text{if } unc \geq \theta \end{cases} \quad (4.1)$$

where θ denotes the threshold of the step mapping function.

Furthermore, we could also assign a high weight (1 in our case) to the background (external) region of the bounding box annotation, as it is an acceptable assumption that no object is outside the bounding box; this can be enforced as guidance to the annotator during drawing of bounding boxes. Subsequently, we combine the background weight map M_{bg} and the uncertainty weight map M_{unc} using the maximum operation. The baseline cross entropy loss is calculated by using Equation (4.2).

$$\mathcal{L}_{ce}(p(x), b(x)) = - \sum_{c=0}^{C-1} b_c(x) \log(p(y = c|x)) \quad (4.2)$$

$b(x)$ represents the bounding box label for input x , and $b_c(x)$ is a binary indicator that is 1 if class c is present within the label for input x and 0 otherwise. $p(y = c|x)$ is the model's estimated probability for class c given input x . The weighted cross entropy loss is then calculated by using Equation (4.3).

$$\mathcal{L}_{UB} = \sum_{x \in \mathcal{X}} [\max(M_{unc}(x), M_{bg}(x)) \times \mathcal{L}_{ce}(p(x), b(x))] \quad (4.3)$$

4.3.3 CRF based spatial constraint loss

By solely utilising the uncertainty-guided cross entropy loss, the model is prone to prioritising the learning of image regions that exhibit similarities across all training examples, potentially disregarding the distinctive characteristics of each individual image. To address this, inspired by the local CRF model proposed by Zhang et al. [Zhang et al., 2020], the segmentation problem can be solved by minimising an energy function containing a unary term and a pairwise term. The unary term is normally obtained for each pixel by a classifier, which refers to the L_{UB} in Eqn. (4.3) in our case. The pairwise term specifies the penalty of assigning labels to the current pixel i and another pixel j at the same time. Using a local CRF model, the pairwise term can be determined by its nearest neighbour pixels, which provides a local smoothness constraint to the model. The pairwise term at pixel i is defined as,

$$\psi_i = \frac{\sum_j \psi_{i,j}}{\sum_j |\mu_{i,j}|}. \quad (4.4)$$

$$\psi_{i,j} = \mu_{i,j} P_i P_j \quad (4.5)$$

$$\mu_{i,j} = \begin{cases} -\exp\left(-\frac{(I_i - I_j)}{2\sigma^2}\right) & (j \in R, \hat{Y}_i = \hat{Y}_j) \\ +\exp\left(-\frac{(I_i - I_j)^2}{2\sigma^2}\right) & (j \in R, \hat{Y}_i \neq \hat{Y}_j) \\ 0 & (j \notin R) \end{cases} \quad (4.6)$$

where P_i and P_j are the predicted confidence scores for pixel i and its neighbour pixel j , respectively. $\mu_{i,j}$ is the weight between them. $\hat{Y}_i$ and $\hat{Y}_j$ are the predicated labels from the model for pixel i and j , respectively. I

here represents the original image intensity. R is the domain of the nearest neighbours for pixel i . Finally, the pairwise CRF term (spatial constraint loss) for the whole image is defined as:

$$\mathcal{L}_{\text{CRF}} = \sum_i \psi_i. \quad (4.7)$$

4.3.4 Total loss function

The computed pairwise term in Eqn. (4.7) is then combined with Eqn. (4.3) to form the total loss function of our proposed segmentation model, expressed as:

$$\mathcal{L}_{\text{UBC}} = \mathcal{L}_{\text{UB}} + \lambda \mathcal{L}_{\text{CRF}} \quad (4.8)$$

where λ is for balancing the unary and pairwise terms.

4.4 Experiments

4.4.1 Datasets

For the experiments in this chapter, we used a public dermatoscopy dataset from the *ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection challenge* dataset [Tschandl et al., 2018] (this dataset was also used in Chapter 3). This dataset consists of 2,594 RGB images of different types of skin lesions with corresponding lesion segmentation masks. For our experiments, the images are resized to 256 x 256 pixels for computational reasons and normalised using zero-mean normalisation and scaled to [0,1] with intensity normalisation method. Each of the bounding box annota-

tions was automatically generated from the original segmentation mask, using a region that covers the entire object with a minimal size. In all of our experiments, 70%, 10%, and 20% of the images are randomly selected for training, validation, and testing, respectively. The model is trained using the ResUnet model described in section 4.3.1. For model deployment only the raw images are required without requiring the bounding box annotations.

4.4.2 Baselines

We conducted a comparison between our proposed method and two existing methods described above: DeepCut (a deep learning-based method) [Rajchl et al., 2016] and GrabCut (a non-learning based method) [Rother et al., 2004]. Additionally, we evaluated two baseline, fully supervised-models: Weak-ResUnet, which used the bounding box as the label, and Full-ResUnet, which used the original pixel-wise annotation as the label. Both baseline models employed the same ResUnet architecture as our proposed method. To quantitatively compare the performance, we used the evaluation metrics Dice Coefficient, Precision, and Recall. In the ResUnet architecture, there are 5 layers in both the encoder and decoder paths. All the models underwent training for 50 epochs until convergence using the Adam optimiser. The learning rate used was 0.0001 with a decreasing step of 10.

For our method, we experimentally evaluated different values for the dropout rate (range between 0 and 0.2), the number of stochastic forward passes M (in equation 3.3, range between 10 and 30), the uncertainty-based weight function threshold θ (in equation 4.1, range between 0.1 and 0.5) using the validation set. Table 4.1 shows the parameter settings.

Parameter	Value
Dropout rate (d)	0.1
Number of stochastic forward passes (M)	20
Uncertainty-based weight function threshold (θ)	0.1
CRF parameter (σ)	1.5
CRF parameter (λ)	1

Table 4.1: Experimentally assessed parameter values for the proposed method.

4.4.3 Result

A comprehensive ablation study was conducted to evaluate the effectiveness of the components of the proposed loss function (UBC loss). The study revealed that the best performance is achieved when these loss elements are used in combination.

Table 4.2 lists the quantitative measurements of all compared methods. We can observe that our proposed method performs better than both DeepCut and GrabCut in terms of Dice and Recall.

Model	Dice Coefficient	Precision	Recall
Full-ResUnet	0.863 $\pm$ 0.16	0.874 $\pm$ 0.17	0.89 $\pm$ 0.16
Weak-ResUnet	0.757 $\pm$ 0.15	0.664 $\pm$ 0.17	0.938$\pm$0.15
GrabCut[Rother et al., 2004]	0.776 $\pm$ 0.28	0.893 $\pm$ 0.27	0.714 $\pm$ 0.28
DeepCut[Rajchl et al., 2016]	0.706 $\pm$ 0.31	0.905$\pm$0.27	0.631 $\pm$ 0.31
UBC (Proposed)	0.823$\pm$0.16	0.821 $\pm$ 0.17	0.866 $\pm$ 0.19

Table 4.2: Skin lesion segmentation performance comparison on test set for baseline Full-ResUnet, Weak-ResUnet, GrabCut, DeepCut and UBC our proposed method.

The distribution of Dice values for test set images, as illustrated in Figure 4.4, provides a clear comparison between our proposed method, DeepCut,

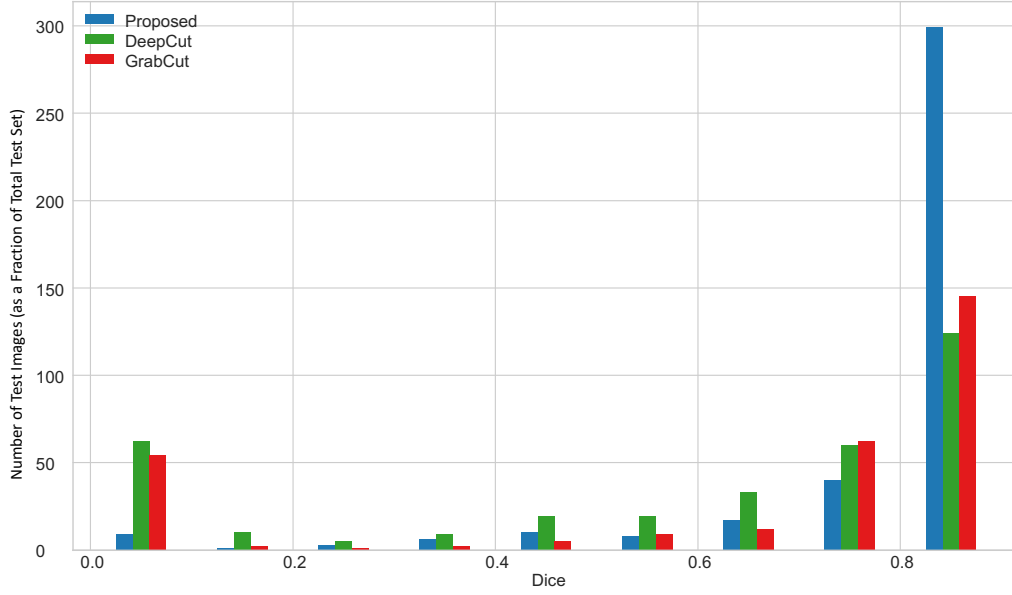


Figure 4.4: Plot comparing the distribution of Prediction Dice Coefficient scores for different methods (Proposed Method, DeepCut, and GrabCut) across the test set. The x-axis represents ranges of Dice Coefficient scores achieved for test set, while the y-axis shows the number of test images falling within each score range, expressed as a fraction of the total number of test images. This visualisation illustrates the frequency with which each method achieves various levels of Dice Coefficient scores on the test images.

and GrabCut. The graph shows an important issue where both GrabCut and DeepCut were unable to segment the lesions accurately in over 50 images out of the total test set, resulting in Dice value of zero. This failure mode is particularly pronounced in the DeepCut method, where the iterative updates of pseudo-masks are prone to errors as discussed in Section 4.2. These errors often accumulate during training, ultimately leading to completely empty predictions.

In an attempt to mitigate this issue, the DeepCut method integrates a dense CRF to refine the predicted pseudo-masks and encourage smoother segmentation results. However, despite this integration, the standalone application of CRF as a post-processing step appears to be insufficient. It fails to counteract the tendency of the pseudo-masks to shrink excessively, which is a critical flaw evident in the method’s design. This drawback

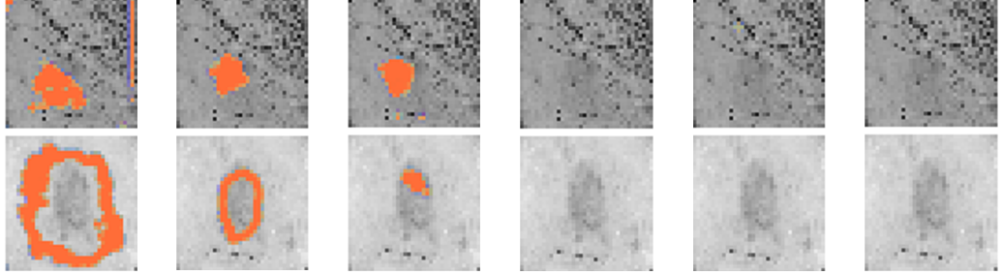


Figure 4.5: Selected DeepCut test results contouring the predicted lesion region over the original image on the models from epoch 1 to epoch 50 with step size of 10 (left to right), showing the vanishing issue of pseudo-masks during training.

is further exemplified in Figure 4.5, which depicts two specific instances from the DeepCut predictions. These examples, taken at various epochs throughout the training process, demonstrate how the pseudo-masks can shrink over time, ultimately leading to the loss of meaningful segmentation results. This comparative analysis underlines the inherent challenges and limitations faced by methods like DeepCut and GrabCut.

To illustrate the advantages of our proposed method in comparison to other existing approaches, we selected three input samples based on their Dice values obtained from a fully supervised model. The prediction results of these samples are visually displayed and compared to the original ground truth in Figure 4.6. In the top row, it is evident that the contrast between the lesion and the surrounding background is quite poor, which posed a significant challenge for both GrabCut and DeepCut, leading to their failure in accurately predicting the lesion. This particular instance underscores the utility of incorporating the local CRF term within our proposed loss function, which effectively enhances the model’s ability to delineate between the lesion and background even under challenging contrast conditions. Moving to the middle row, it is observed that GrabCut tends to over-segment the lesion area. The bottom row presents an image that is arguably one of the most challenging in the entire dataset due to the lesion’s dual-textured

nature. In this complex scenario, among the weakly supervised methods evaluated, GrabCut managed to achieve a comparatively better prediction for this image, while DeepCut yielded the least accurate result, further highlighting the strengths and weaknesses of each method under varying conditions.

4.5 Discussion

The observed progression of our proposed UBC method from epoch 1 to epoch 20 on sample images from the ISIC2018 dataset, depicted in Figure 4.7, provides deeper insights into the dynamic nature of our approach during training. Initially, the model utilises bounding box annotations (c) to guide the segmentation, and as training progresses, it continuously refines its predictions (d), attempting progressively towards the ground truth (b). This refinement is driven by the innovative incorporation of customised weighted cross-entropy and CRF terms in our loss function, designed to employ both the epistemic uncertainty and the spatial relationships present in the data.

The uncertainty heatmap (e), transitioning from high (red) to low (dark blue) uncertainty, reflects the model’s progress toward increased confidence in identifying lesion regions. Initially exhibiting high uncertainty even in background regions, the dynamic changes in the heatmap are consistent with the learning phase, where the model is still getting familiar with the data. As training advances, the influence of the background constraint in our loss function becomes more pronounced, noticeably enhancing the model confidence and accuracy, within the background regions, as denoted by the dark blue in the heatmap.

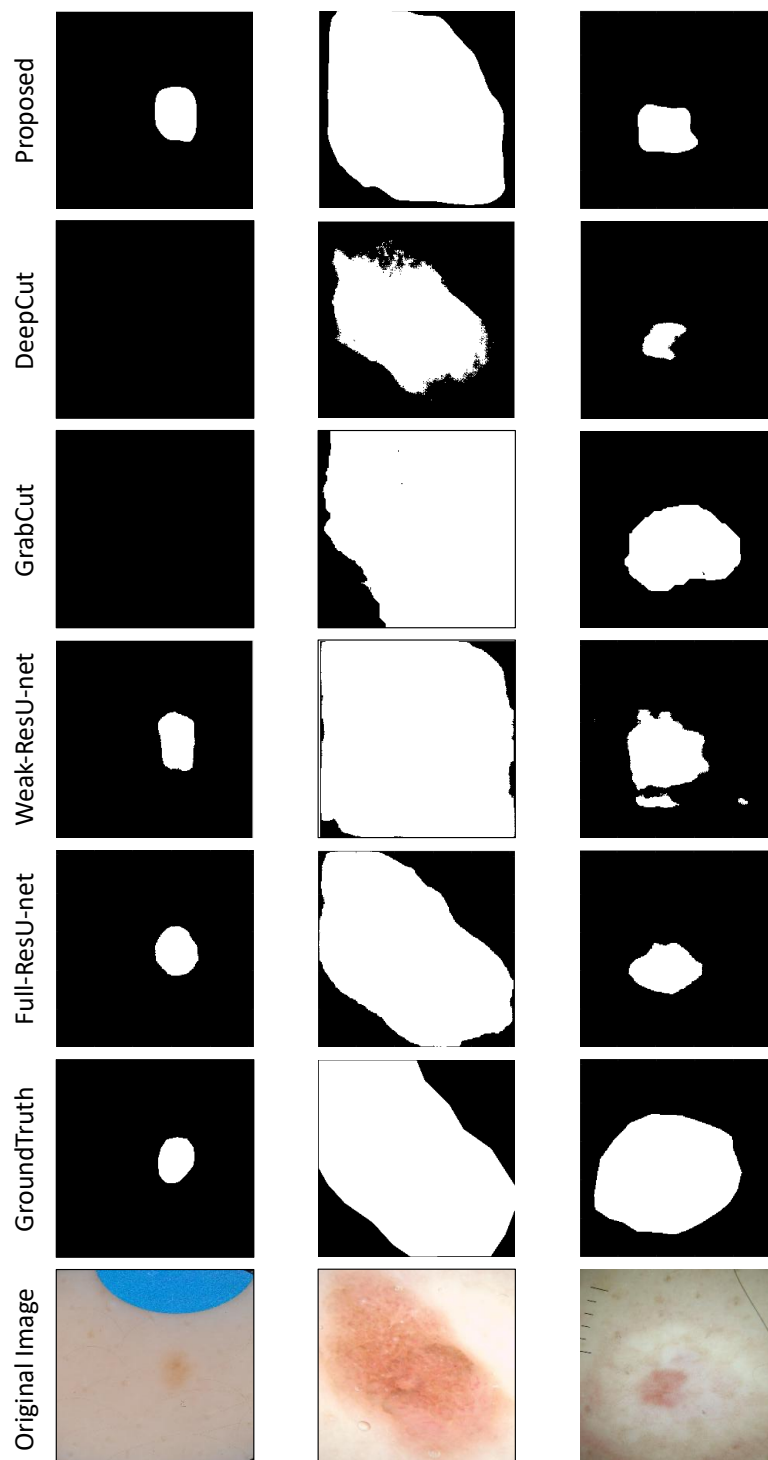


Figure 4.6: Example skin lesion segmentation results for all compared methods: from left to right (a) original image, (b) ground truth, (c) Full-ResUnet, (d) Weak-ResUnet, (e) GrabCut, (f) DeepCut, (g) Proposed method.

The uncertainty heatmap also effectively captures persistently challenging regions where the model struggles, highlighting areas that require further refinement. This consistent correlation between the heatmap depiction of uncertainty and the model’s actual performance provides a valuable tool for understanding and analysing the model behaviour and limitations at each training stage. It is essential to recognise and evaluate the limitations of our approach in addressing complex segmentation tasks, even though the observed improvements are promising.

In analysing the progression of our UBC method for the two distinct samples from the ISIC2018 dataset shown in Figure 4.7, we observe unique behaviours attributed to the contrast of foreground and background intensity in each case. For the *sample 1* shown in the figure, with a lesion that is distinct from the background, the tailored loss successfully aids the model in maintaining the accurate shape of the lesion. The clear intensity differentiation between the lesion and the background allows the CRF to reinforce the model’s confidence in the lesion’s boundaries, effectively countering the background constraint’s tendency to shrink the foreground. This results in a more precise segmentation, avoiding the incorrect inclusion of background regions, as marked by the yellow dotted circle.

Although we observed that CRF could be very helpful in preventing the model from generating empty predictions, it could sometimes lead to incorrect segmenting of some inner corner regions of the bounding box as part of the lesion. For example, for the *sample 2* in the figure, where the lesion shares a similar intensity with the surrounding background, the CRF term in the loss function led to inaccurate segmentation. Despite successive epochs of training, the model sometimes misinterprets the lesion boundaries due to the subtle intensity differences, leading to over-segmentation particularly near the ambiguous border regions, as indicated by the yellow

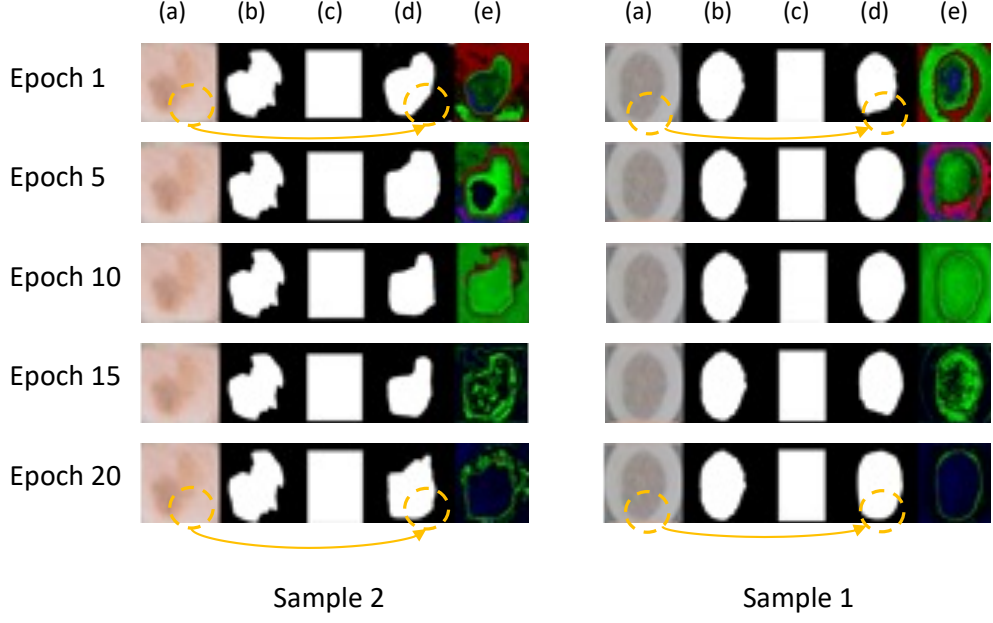


Figure 4.7: The proposed UBC method progress during training from epoch 1 to epoch 20 on 2 sample images from the ISIC2018 dataset: (a) Original Image; (b) Ground Truth; (c) Bounding Box Annotation; (d) Prediction Result; and (e) Uncertainty Heatmap with red as the highest uncertainty value and dark blue as the lowest uncertainty value.

dotted circles. This illustrates the challenge of delineating regions where the foreground and background blend seamlessly.

The different scenarios mentioned emphasise the importance of the CRF in improving the segmentation process. However, they also bring attention to a limitation of our proposed tailored loss function when dealing with samples that have lesions with subtle intensity variation compared to the background.

The high recall observed in our model underscores its effectiveness in identifying most relevant lesion areas across the dataset, reflecting its ability to minimise false negatives in comparison to DeepCut and GrabCut. This characteristic is particularly crucial in medical imaging contexts, where the cost of missing a lesion can be significantly higher than erroneously identifying non-lesion regions as lesions.

However, the lower precision associated with our model in comparison with DeepCut and GrabCut highlights a tendency towards over-segmentation, where the model incorrectly segments some non-lesion pixels as part of the lesion. This behaviour is exemplified in our observations of the occasional incorrect segmentation within inner corner regions of the bounding box, attributed to the CRF component of our loss function.

In medical imaging, precision is critical for ensuring that treatments are accurately targeted and do not harm healthy tissues. Considering the risks of low-precision segmentation becomes even more important in cases where treatments are directly guided by segmentation results, such as in radiation therapy planning for cancer treatment. For instance, studies have shown that segmentation performance is highly correlated with tumour control and radiation toxicity. Low-precision for example, could lead to higher than intended doses to healthy tissue in radiation therapy [Harrison et al., 2022]. Moreover, in the diagnosis phase, low-precision segmentation might contribute to false positives, leading to misdiagnosis and inappropriate clinical decisions.

To mitigate the risks associated with low precision in segmentation, improvements to the proposed approach could involve refining the loss function to incorporate additional penalties for false positives to help balance the trade-off between recall and precision. For instance, integrating a precision-focused regularisation term could reduce the tendency of the model to over-segment. Given the focus on uncertainty in the proposed model, further enhancements could involve tailoring the loss function to reflect higher uncertainties in the uncertainty map at inference for regions that the model normally oversegments to help with decision making.

These insights into our model’s performance, characterised by high recall

but lower precision, suggest a need for further refinement of our approach, especially in enhancing its specificity without significantly compromising its sensitivity. Future versions of our model could benefit from exploring adjustments to the loss function that more effectively balance the trade-off between recall and precision, potentially through incorporating more discriminative features.

4.6 Conclusion

In this chapter, we have presented a tailored loss function for weakly supervised skin lesion segmentation using bounding box annotations. Our approach, illustrated in Figure 4.3, integrates a novel loss function comprising both weighted cross entropy and CRF terms. The weighted cross entropy is uniquely informed by an epistemic uncertainty estimation, which, alongside a bounding box background constraint, refines the focus of the model training. The CRF term further regularises the predicted labels based on the original image intensity, independent of the bounding box annotation. We evaluated the performance of our method using a dermoscopic dataset containing challenging skin lesions with various sizes, textures and contrasts.

The efficacy of our proposed method was demonstrated through comparisons with state-of-the-art methods DeepCut and GrabCut, as well as baseline fully supervised and weakly supervised models. The results, detailed in quantitative measurements of Dice Coefficient, Precision, and Recall, show our method outperforms GrabCut and DeepCut in terms of dice and recall. However, it is crucial to acknowledge certain limitations inherent to our approach.

While our method shows promising results, it is not without drawbacks. One challenge arises when the CRF component of our loss function could sometimes mistakenly lead the model to incorporating background regions inside the bounding box into the lesion area. This is particularly noticeable in the ambiguous border regions, where the model sometimes misinterprets the corner regions inside the bounding box. Based on these observations, future work should aim at refining the CRF implementation to better handle subtle intensity variations.

Chapter 5

Understanding Multiple Sclerosis: Deep Convolutional Neural Networks for Multiple Sclerosis Lesion Segmentation, and Associated Challenges

5.1 Context

In diagnosing MS, a complex condition of the central nervous system, MRI has emerged as a critical diagnostic tool. Despite advancements in imaging techniques, such as T2 FLAIR and T1-weighted imaging, challenges persist in accurately segmenting MS lesions. This accurate segmentation is crucial for leveraging the Central Vein Sign (CVS) [Maggi et al., 2018] as a diagnostic marker, highlighting the pressing need for robust segmentation methodologies to enhance diagnostic accuracy and patient management in

MS.

The variability and subtlety of MS lesions make them particularly difficult to detect and segment accurately. Given its significant implications for patient care and treatment, MS represents a critical application area for testing and refining some of the methodologies proposed in previous chapters. As such, this chapter is dedicated to the segmentation of MS lesions. Therefore, we first introduce and review the specific literature related to MS lesion segmentation, acknowledging its absence in the main literature review due to the specialised nature of this discussion.

The contributions in this chapter are as follows:

- **Comprehensive Exploration of MS Data for Segmentation:** Our work encompasses a thorough examination of MS data preparation, pre-processing techniques, and the evaluation of DCNN models, aiming to optimise MS brain lesion segmentation.
- **Evaluation of Preprocessing Approaches:** For the data preparation for model training, we assessed two preprocessing methods, for brain tissue extraction, through visual inspection and segmentation performance analysis.
- **Custom Loss Functions for Enhanced Robustness:** We showcased the importance of incorporating model uncertainty feedback and data uncertainty as regularisation to enhance segmentation performance, especially in scenarios involving unseen data from various sites and scanners.

5.2 Introduction

Medical image segmentation is one of the most critical steps in many clinical applications. In brain MRI analysis, image segmentation is commonly used for analysing brain changes and for characterising pathological regions. The diversity of image processing applications has led to the development of various segmentation techniques with different accuracies and degrees of complexity [Despotović et al., 2015]. In radiology, lesions are visually identified, but manual segmentation specifically in brain lesions of MS is a time-consuming process and subject to human errors. Consequently, automated MS lesion segmentation techniques can be beneficial for an accurate and swift diagnosis [Danelakis et al., 2018].

5.2.1 Multiple Sclerosis

Multiple Sclerosis, an inflammatory demyelinating disease of the central nervous system, is characterised by lesions in the brain and spinal cord [Dworkin et al., 2018, Tallantyre et al., 2011]. Dissemination of a multifocal demyelinating lesion in space and time is a key component of the current diagnostic criteria for the disease. Thus, detecting and segmenting MS lesions is an indispensable step in characterising the disease, and also in calculating and interpreting more specialised metrics of damage [Aktas et al., 2018, Brownlee et al., 2017, Tallantyre et al., 2011, Danelakis et al., 2018]. When a patient presents symptoms that could be evidence of MS, an MRI is highly recommended for the reason that an abnormal brain MRI is present in almost all MS patients and in more than 80% of patients with clinically isolated syndrome who go on to develop MS [Brownlee et al., 2017]. MRI is highly sensitive in detecting MS plaques and has been ac-

cepted as a key clinical tool for diagnosing MS, determining the progression and the activity of the disease and assessing the success of medical treatments [Danelakis et al., 2018].

T2 FLAIR, along with T1-weighted magnetisation-prepared rapid gradient-echo imaging (T1) as an alternative confirmatory sequence, is recommended as a primary identification tool for Periventricular, Juxtacortical/Cortical, and Infratentorial MS lesion types [Filippi et al., 2019].

T1 MRI modality, through 3D MPRAGE, is a three-dimensional imaging technique applicable for MR imaging throughout the body [Mugler III and Brookeman, 1990]. It is particularly adept at capturing high-quality 3D image sets, as evidenced in its ability to produce images of the brain displaying excellent grey and white matter contrast [Mugler III and Brookeman, 1990]. T1 structural imaging is essential for obtaining crucial data on the volumes and structures of brain tissues. In addition, it is crucial for computing cross-subject and cross-modality alignments, which are necessary for analysing all other brain modalities [Alfaro-Almagro et al., 2018]. The 3D MPRAGE acquisition method begins with a Magnetisation Preparation (MP) period, in conjunction with the snapshot FLASH [Haase, 1990] technique. An inversion pulse [Bydder and Young, 1985] followed by a delay is typical for yielding T1-dependent contrast, crucial for differentiating between various tissues and aiding in precise structural assessments. After the longitudinal magnetisation is encoded with the desired contrast during the MP period, a RApid Gradient-Echo (RAGE) [Hargreaves, 2012] sequence is used to sample the prepared magnetisation, ensuring detailed and accurate imaging essential for further medical analysis and treatment planning.

T2 FLAIR imaging is a specialised MRI technique that provides high

contrast images by suppressing the signal from fluids, particularly CerebroSpinal Fluid (CSF) [Alfaro-Almagro et al., 2018]. This suppression makes T2 FLAIR exceptionally useful for depicting tissue alterations typically associated with various pathologies, such as white matter lesions found in diseases such as MS, infarction, subarachnoid haemorrhage, and head injuries. The contrast in T2 FLAIR images is dominated by signal decay from interactions between water molecules, known as T2 relaxation times, which allows detailed visualisation of brain tissues and structures [Alfaro-Almagro et al., 2018]. The suppression of CSF makes T2 FLAIR images particularly useful in highlighting subtle changes at the periphery of the hemispheres and in the periventricular region close to CSF, offering an advantage over other sequences such as proton density or standard T2-weighted images where CSF appears bright [Okuda et al., 1999].

The usefulness of FLAIR sequences extends to a wide range of central nervous system diseases. However, FLAIR sequences are usually obtained in addition to T1 and T2-weighted images, as they are expected to provide additional useful information, particularly in the detection and characterisation of brain lesions [Okuda et al., 1999]. Despite their many benefits, FLAIR sequences are not without limitations. Incomplete nullification of CSF signals due to CSF inflow effects can produce imaging artifacts, especially in areas with prominent CSF pulsatility, which could contribute to failure in detecting lesions in certain regions such as the brain stem [Okuda et al., 1999]. Moreover, the long acquisition time required for conventional FLAIR sequences and the insufficient lesion contrast in some areas such as the basal ganglia and posterior fossa are recognised challenges [Okuda et al., 1999]. Despite these drawbacks, the overall advantages of FLAIR imaging make it an invaluable tool in the diagnosis and study of various neurological conditions. Figure 5.1a depicts a brain MRI of a patient di-

agnosed with MS; however, the particular slice chosen for this image does not show any MS lesions. Figure 5.1b on the other hand, illustrates a slice from a different MS patient’s brain MRI, where the MS lesion is clearly visible in both the coronal and sagittal views presented in the figure.

The misdiagnosis of MS has been shown in previous studies as an issue in clinical practice and the need for an accurate diagnosis is important due to the increasing number of immunosuppressive medications used in patients with MS [Brownlee et al., 2017, Aktas et al., 2018, Solomon et al., 2016, Dworkin et al., 2018, Samaraweera et al., 2017]. Misinterpretation of brain MRI has been considered one of the main causes of misdiagnosis. Consequently, a robust imaging biomarker would be beneficial for patients and clinicians [Samaraweera et al., 2017, Solomon et al., 2016]. As a means of discriminating MS lesions from white matter abnormalities caused by other diseases, the identification of a vein crossing the center of a lesion has been proposed as a diagnostic tool due to the fact that inflammatory demyelination in the MS white matter is perivenular (around the vein) [Dworkin et al., 2018, Tallantyre et al., 2009, Sati et al., 2016]. Figure 5.2 shows the perivenular lesion in 3T Flair* MRI of a subject with relapsing and remitting multiple sclerosis.

Strong evidence has been shown in research studies that higher proportions of MS lesions show the CVS in comparison to lesions caused by mimics of MS [Sati et al., 2016, Mistry et al., 2013, Dworkin et al., 2018, Maggi et al., 2018]. However, there are concerns regarding the CVS subjective assessment variability and the time that is required to decide on CVS in every lesion per patient. These concerns could limit the practicality of the clinical application of CVS.

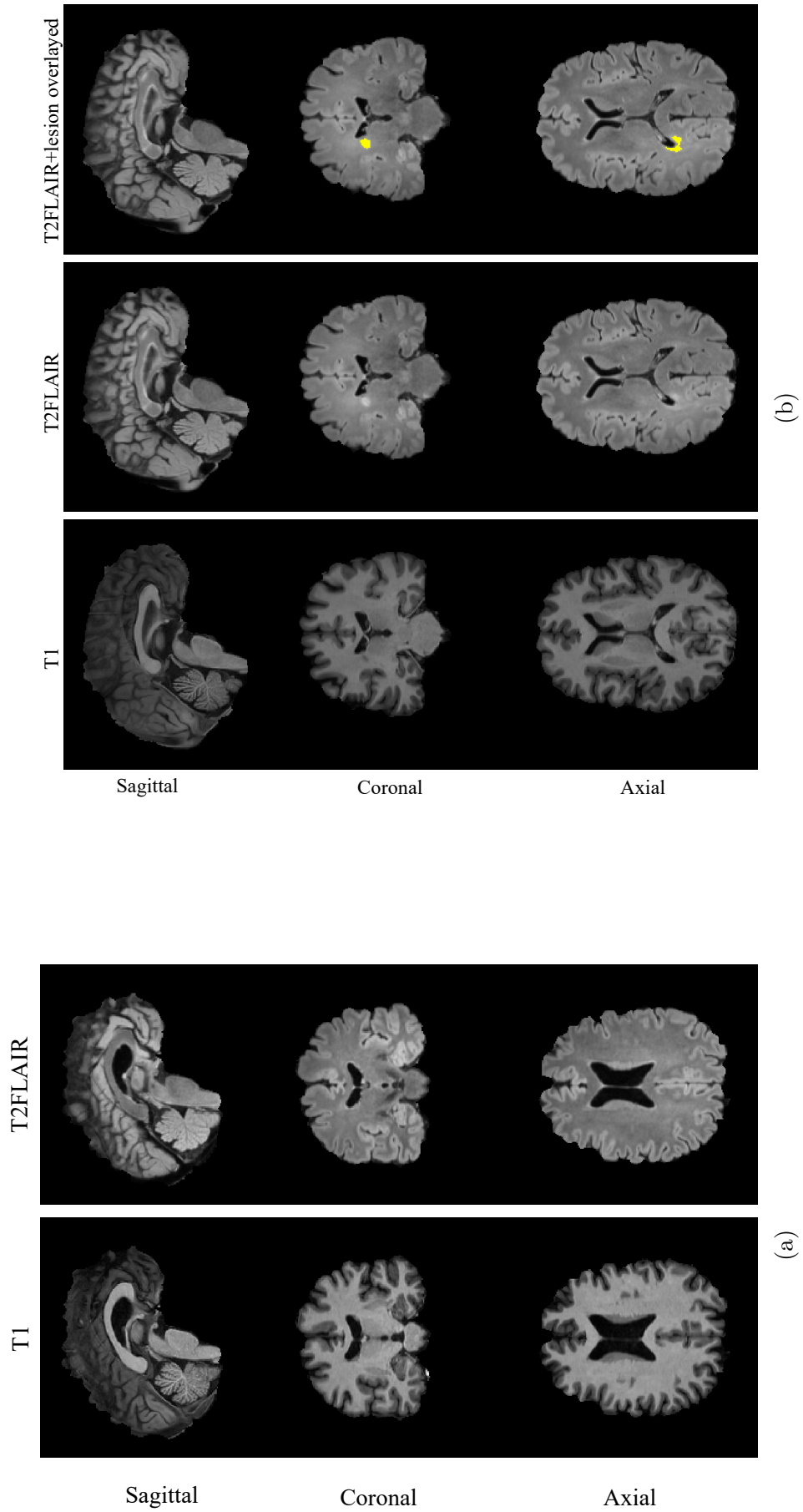


Figure 5.1: (a) Presents T1 and T2 FLAIR brain MRI scans of an MS patient, showcasing a slice where no lesions are detected. (b) shows T1 and T2 FLAIR brain MRI scans of another MS patient, highlighting a slice where a lesion is clearly visible in yellow in both the coronal and axial views.

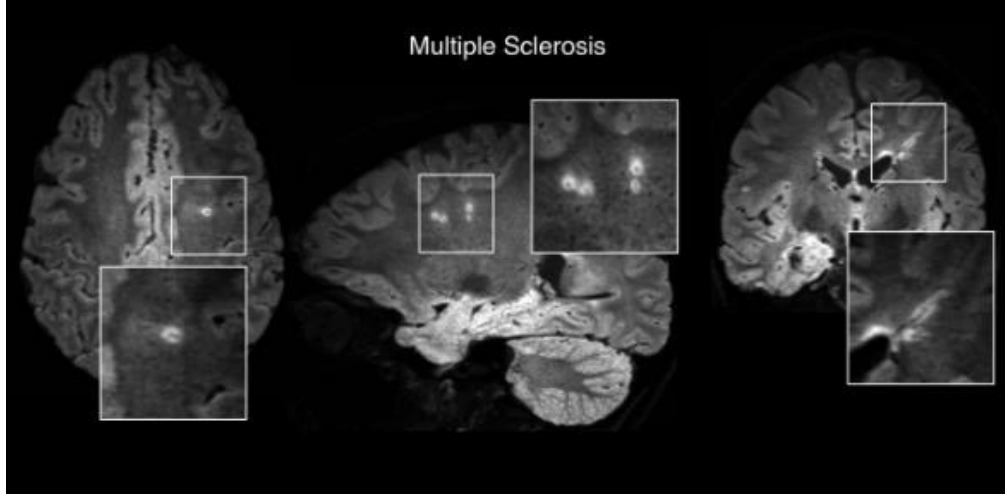


Figure 5.2: Axial, sagittal, and coronal 3T FLAIR* images of relapsing-remitting multiple sclerosis; showing MS lesions with central vein sign [Maggi et al., 2018]

Although some studies suggested methods to limit the number of lesions to be examined, by generating rules for fast allocation of disease category or restricting the candidate lesions [Solomon et al., 2018, Mistry et al., 2016], these approaches could not be as successful as the assessment of CVS in all lesions per patient. These approaches may also increase the variability of assessments [Dworkin et al., 2018]. The first automated technique for MS detection using CVS in MRI proposed a six-step approach which requires T1-weighted volume, T2-weighted FLAIR volume, and T2*- weighted segmented Echo-Planar Imaging (EPI) volume.

In the first step, by T2*-EPI, a map of veins is created and registered to T1 volume. In the second and third steps, using T1 and T2 FLAIR volumes, the white matter lesions are segmented and the lesion boundaries are determined by removing the ambiguous boundary voxels. The fourth step is to remove the periventricular lesions from the candidate lesions. The fifth step is to determine whether identified veins occur in the centre of a given lesion to a greater degree than expected by chance. The last step is to achieve more stable biomarkers by weighing the contribution of each lesion to the

biomarker by the amount of noise in the voxels of the lesion on the T2*-EPI volume. Although the resulting automated biomarkers of this study had strong diagnostic ability, the automated decision of every lesion in a scan is not yet as accurate as manual categorisation of every lesion in a scan. More studies are essential for this approach to be feasible for use in the clinic. Moreover, due to the noise in T2*-EPI, 9 subjects were excluded from the study, implying the lack of robustness of the proposed method. Furthermore, biomarker values were found to be lower than previously reported CVS proportions for MS patients and higher than previously reported CVS proportions for those without MS. This could arise from the poor performance of the lesion segmentation step [Dworkin et al., 2018]. Hence, it is evident that accurately segmenting lesions is crucial in differentiating an MS brain from other diseases that resemble MS. This step is essential in reducing misdiagnosis and avoiding unnecessary medication for patients before addressing CVS biomarker detection. In this chapter, our focus is on the white matter lesion segmentation using supervised methodologies.

5.2.2 Multiple Sclerosis Supervised Lesion Segmentation

Supervised MS lesion segmentation methods undergo a training process in order to learn the definition of lesions from the data that are usually previously manually segmented. The training is based on the features extracted by the methods. The MS supervised lesion segmentation methods can be categorised into four subcategories of atlas-based, data-driven, feature-based and statistical approaches [Danelakis et al., 2018].

Atlas-based methods

The segmentation of medical images often needs inclusion of *a priori* information. Brain map (atlas) can provide statistical or topological information. A statistical atlas gives the prior probability of each voxel of a specific tissue, and a topological atlas encodes a certain topology for each structure and group of structures. Techniques that use atlas information require the application of a registration process to fit the atlas to the input image [Danelakis et al., 2018]. Generally, the atlas is transformed to match the input image, and to segment the structures. Another approach is to implement a non-rigid registration [Dawant, 2002] of a single atlas to the image with the appropriate structures segmented by mapping the segmentation labels. Other proposed approaches combine statistical atlas and non-rigid registration [Bazin and Pham, 2007]. In [Jesson and Arbel, 2015], a hierarchical framework for the segmentation of healthy tissues and MS lesions was proposed. The study used ISBI 2015 Challenge data [Carass et al., 2017] and in addition to the preprocessing provided by the challenge, intensity normalisation and a block wise non-local means denoising filter were done [Coupé et al., 2008]. Their argument is that by modelling healthy structures in the presence of pathology, lesion segmentation can be enhanced. In the proposed framework, first atlases were built for lesions and a wide variety of healthy tissues off-line during training. At the second stage, Multi-Atlas Label Fusion (MALF) was performed from the atlases to patient images to use as spatial priors for a voxel-based MRF framework. Finally, a Regional Random Forest classifier refined candidate lesions segmentations. Another study proposed Lesion-TOADS method which is an extension of the Topology-Preserving Anatomical Segmentation (TOADS) [Bazin and Pham, 2008] algorithm. Lesion-TOADS is formulated based on the fact that topological outliers, such as lesions, can

be addressed in a topology-preserving framework when they are grouped together with the underlying tissues. The proposed method models the lesions while maintaining the constraints provided by topological and statistical atlases. Before segmentation, the atlases are rigidly registered to the MR image of interest. Then, MS lesions can be modeled as a separate cluster in the TOADS intensity atlas-based segmentation algorithm with the desirable topological properties. As significant amount of false positives often occurs in a tissue classification or clustering technique due to the intensity profile similarity of lesions to other structures in the brain; a post-processing step for reducing false positives is applied as the final step [Shiee et al., 2010]. The primary dissimilarity is that Shiee et al. focuses on topology-preserving segmentation and post-processing to mitigate false positives [Shiee et al., 2010], while Jesson et al. emphasize hierarchical processing by incorporating MALF [Jesson and Arbel, 2015].

Data-Driven Methods

These methods utilise data-based approaches such as thresholding [Sezgin and Sankur, 2004], spatial analysis, intensity analysis, morphology analysis, topology analysis or region growing [Adams and Bischof, 1994]. In [Prados et al., 2015], the Optimised Patch Match Label fusion method [Ta et al., 2014] is used to locate pathological regions by using a template library containing multimodal images with manually segmented MS lesions. Patch Match estimates the location of the lesions in the target image by matching patches between the template library images and target images. As the proposed method has a patch matching methodology, it is essential that the template database comprises of a wide variety of lesion sizes and lesions with various locations which could limit the practicality of this method. In [Guizard et al., 2015], a Rotation-invariant Multi-contrast Non-

Local Means Segmentation (RMNMS) method is proposed for MS lesion segmentation. The preprocessing steps include: Non-Local Means (NLM) image denoising [Coupé et al., 2008], intensity non-uniformity correction [Sled et al., 1998], linear intensity normalisation of the image histogram, linear registration of each T1W image, Rigid registration of the T2W and FLAIR to the T1W image, and brain extraction [Eskildsen et al., 2012]. The NLM algorithm in this study uses a rotation-invariant distance metric, instead of the L2-norm metric. The results of RMNMS, with multi-contrast and rotation-invariant patch distance, shows that the non-local method is able to detect structures that vary in size, shape and location such as MS lesions. However, the proposed method relies on a set of dataset containing a variety of MS lesions with different properties. Thus, it requires a larger number of training subjects to capture the variability of the structure to be segmented where spatial distribution, shape and size of MS lesions are not consistent. The similarity of Prados et al. and Guizard et al is the reliance on template-based approaches which could limit their efficacy.

Feature-based methods

These methods are based on the features that are extracted from the input data containing meaningful information. These features, can then be used for further processing. In [Jog et al., 2015], as the first step the N4 [Tustison et al., 2010] (modified nonparametric nonuniform intensity normalisation) for bias correction, rigid registration and skull stripping is performed. The features extracted are local intensity information and global context information for a voxel and are fed into multi-output decision trees. As the last step the output image is smoothed using a Gaussian filter to reduce false positives [Danelakis et al., 2018]. In another approach, segmentation is used to obtain a hierarchical decomposition of a multichannel MR scans.

Unlike other approaches that use voxel-wise analysis, the proposed system uses regional properties which are important for characterising abnormal brain structures. A set of features are extracted including intensity, shape, location, neighborhood relations, and anatomical context of segments; and a supervised decision forest classifier is trained based on the obtained features. The main errors of this approach, however, are due to the false positive rates related to dirty-appearing white matter (brain regions of intermediate signal intensity between those of focal lesions and normal-appearing white matter) rather than focal lesions [Akselrod-Ballin et al., 2009].

In [Strumia et al., 2016], a 3D MS lesion segmentation method based on an adaptive geometric brain model is proposed. The topological features of the lesions and brain tissues are modeled in order to constrain the lesion segmentation to the White Matter (WM). As a result, the method is independent of an MRI atlas. The proposed geometric model of the brain is less sensitive to spatial variabilities of the brain’s anatomy. An energy is defined which models the appearance of the tissues and the topological structure of the brain including the ordering of the different tissue layers and the connectivity of the grey matter. One limitation of the proposed method is that the geometric model is not completely correct, as the entire adjacencies for the WM are not modeled.

Another study used Random Forest (RF) to be trained on the intensity-based features extracted from the MR scans. The preprocessing is done using a learning based intensity standardisation method. At the post-processing stage, the output of the RF classifier is thresholded with a bias in favour of the lesion class. Finally, single unconnected lesion voxels are removed, holes in binary lesion objects are closed and a single-iteration closing operation with a 3D square-connected component is applied [Maier

and Handels, 2015].

In [Steenwijk et al., 2013], signal intensity, MNI-normalised spatial coordinates x , y and z , and tissue type probabilities p_{CSF} , p_{GM} , and p_{WM} , obtained by Multi-Atlas segmentation method, were utilised as features in k -Nearest Neighbour classifier. The features are normalised based on robust range normalisation and histogram matching. The classifier is used to decide whether a voxel is a lesion or not. Finally, a post-processing step is implemented to reduce false positives based on the lesion volumes. One of the limitations of the proposed method is that a new model needs to be trained when applied to data originating from other scanners or other acquisition protocols. In [Mahbod et al., 2016], four steps of pre-processing including histogram matching, bias field correction, extremely low intensity values removal and normalisation are applied to the MR scans. Then, intensity-based and spatial-based features are extracted and imported to a multi-layer perceptron classifier for a per voxel segmentation. In [Muschelli et al., 2016], the MRI images were corrected, using the N4 inhomogeneity correction. Then, features are extracted such as the normalised intensity of the voxels, their neighborhood and their contralateral differences. Finally, an RF classifier is trained in order to provide the probability that a voxel belongs to a lesion.

In [Vera-Olmos et al., 2016], a learning-based intensity standardisation method is used at the preprocessing stage. Then, a number of intensity-based and location-based features are extracted and are fed into a random forest classifier. This classifier derives an initial lesion mask and a lesion probability mask. A Markov Random Field (MRF) post processing algorithm is applied to make lesions grow through probable neighborhoods using the exploited mask.

While all methods focus on feature extraction from MR images, they differ in their preprocessing steps, the features they extract, and their choice of classifiers and post-processing techniques. These variations impact their effectiveness in handling false positives and adapting to different imaging protocols.

Statistical methodologies

These methods utilise probability density function estimations for probabilistic segmentation. In [Ghafoorian and Platel, 2015], a DCNN with five layers is used to create a statistical, voxel-based classifier. As the preprocessing step, the image intensity is normalised. Then the images are fed into the model as patches and the CNN learns if the central voxel of a given region of interest is a lesion or not. Stochastic gradient descent is used for the network optimisation. In [McKinley et al., 2016], a DCNN for MS lesion segmentation is proposed. To ensure that low-level and high-level features are utilised for the prediction, the proposed network is a combination of an encoder-decoder network and a fully convolutional network with 18 layers. Similarly, another method was proposed in [Brosch et al., 2016], for automatic segmentation of MS lesions based on deep convolutional encoder networks with shortcut connections. The joint training of the two pathways of the networks for feature extraction and prediction allows for the automatic learning of features at different scales that are tuned for a given combination of image types and segmentation task. Furthermore, a new loss function is proposed based on the combination of sensitivity and specificity to address the unbalanced problem of MS MRI scans.

In [Valverde et al., 2016], a CNN network is proposed with two dense convolution layers to probabilistically classify the voxels. The network is

trained using 3D patches to obtain a higher number of training samples whilst reducing the amount of parameters to optimise as network weights. At the preprocessing stage, NL-Means based denoising, rigid registration, skull stripping, bias correction using the N4 algorithm and normalisation is performed on the MRI scans. As an extension of their previous work, a cascade of two 3D patch-wise CNNs is proposed. The first network is trained to be more sensitive detecting the candidate lesion voxels and the second network to reduce the number of misclassified voxels resulting from the first network. The previous proposed CNN is improved to a more sensitive cascaded 3D CNN which tends to learn well from small labeled datasets. This can be beneficial in practice considering the difficulty of manual label annotations and the amount of available unlabelled MRI data [Valverde et al., 2017].

In [Subbanna et al., 2015], an Iterative Multilevel probabilistic Graphical model for the detection and segmentation of MS lesions is proposed. The model has two MRF levels. At the bottom level, a regular grid voxel-based MRF identifies potential lesion voxels, as well as other tissue classes, using local and neighbourhood intensities and class priors. Contiguous voxels of a particular tissue type are grouped into regions. A higher, non-lattice MRF is then constructed, in which each node corresponds to a region, and edges are defined based on neighbourhood relationships between regions. The goal of this MRF is to evaluate the probability of candidate lesions, based on group intensity, texture and neighbourhood regions. The inferred information is then propagated to the voxel-level MRF which performs voxel wise classification. The strength of this model is in detecting very small lesions.

In [Roy et al., 2018], the proposed method contains two convolutional pathways. The first pathway consists of multiple parallel convolutions for mul-

multiple MR modalities. In the second pathway, the outputs of the first one are concatenated and another set of convolutional layers is applied. Furthermore, a preprocessing step is applied including registration, skull stripping and N4 intensity inhomogeneity. In [Kang et al., 2020], a 3D Attention Context U-Net (ACU-Net) was introduced, an end-to-end framework designed for the segmentation of MS lesions with MRI scans. Recognising the challenges posed by MS lesions, which are diffuse, multifocal, and often affect peripheral brain structures, ACU-Net integrates a 3D context-guided module in both the encoder and decoder stages of the traditional 3D U-Net architecture. This module is tasked with expanding the perception field and enhancing the contextual information guiding the segmentation process. Additionally, the framework employs a 3D spatial attention block within the skip connection stages to strengthen the feature representation, focusing on spatial details and improving the overall segmentation of lesions. The effectiveness of ACU-Net was validated on the ISBI 2015 longitudinal MS lesion segmentation challenge dataset.

In [Zhang et al., 2021], Zhang et al. developed a deep learning model based on the 3D U-Net architecture, named ALL-Net, aimed at enhancing the segmentation of MS lesions with high speed and precision, utilising sequences from T2 FLAIR, T1-weighted, and T2-weighted MRI. This model incorporates two novel components: distance transformation mapping to leverage anatomical information, and a lesion-specific loss function to boost the detection of smaller lesions. The effectiveness of ALL-Net was validated through tests on the ISBI 2015 dataset and a more extensive Cornell MS dataset. ALL-Net’s novel approach includes integrating voxel position learning for anatomical detail encoding and lesion-wise learning for improved lesion identification, addressing the challenges of ambiguous image contrast and bias due to lesion location. It effectively utilises anatomical

location information and a lesion-wise loss function, which treats lesions of varying sizes with equal importance, facilitating a shift from voxel-wise to lesion-wise detection. The study also demonstrated the advantage of 3D convolution in capturing richer contextual information for better network training, emphasising the importance of a large dataset to mitigate overfitting risks.

In [Kamraoui et al., 2022], Deep Lesion Brain (DLB), a deep learning framework was proposed that is specifically designed to address domain shift challenges in the segmentation of MS lesions, leveraging a 3D CNN approach. This method has three primary innovations: the deployment of a broad ensemble of compact 3D CNNs for resilient predictions, the application of hierarchical specialisation learning for an effective blend of global and local image feature extraction, and the introduction of a novel image quality data augmentation strategy to minimise training data specificity reliance. DLB utilises only T1-weighted and T2 FLAIR MRI modalities, ensuring compatibility with a broad array of datasets. Through an ablation study, the contributions of each element to the model’s segmentation accuracy and its ability to generalise across domains were highlighted. The motivation behind DLB emerges from the observed limitations of existing CNN-based segmentation methods in their application to clinical data from unseen datasets, despite their proven efficacy in controlled conditions.

This category of methods use statistical methodologies and deep learning approaches, and their main difference are normally in architecture, pre-processing, and strategies for handling class imbalances and domain shift issue. These differences impact their effectiveness in lesion segmentation across various MRI datasets.

The MS lesion segmentation categories, with their key characteristics and

associated limitations are summarised in Table 5.1.

Category	Characteristics	Limitations
Atlas-based	Anatomical atlases, registration	Registration Dependence: Requires accurate registration methods; Atlas Quality: Performance depends on atlas quality; Generalisation: May not adapt well to different datasets.
Data-driven	Data analysis, patch matching, spatial analysis	Data Dependency: Needs extensive, diverse datasets; Adaptability: Struggles with new or varied datasets; Preprocessing Complexity: Complex steps which can introduce errors.
Feature-based	Image features utilisation	False Positives: Issues with ambiguous regions; Preprocessing Impact: Sensitive to preprocessing quality.
Statistical	Probabilistic models for estimating probability density function	Data Dependence: Relies on data quality and quantity; Generalisation: May not generalise well without adaptation; Overfitting: Risk with limited or non-diverse data; Preprocessing Sensitivity: Affected by preprocessing quality.

Table 5.1: Overview of MS lesion segmentation methods, highlighting their key characteristics and associated limitations.

In the rest of this chapter, we focus on this particular challenge of segmenting brain lesions associated with MS. We apply here some of our existing knowledge and ideas onto the challenging and limited-data domain of MS medical imaging, explore the intricacies of preparing MS images for seg-

mentation, and conduct experiments with three in-house MS datasets. Our exploration involves comparing two preprocessing approaches and DCNN models. Specifically, our objective is to assess the effectiveness of our tailored loss functions (UCF, ELF, UCF+AL and ELF+AL introduced in Chapter 3) in the context of Fully supervised MS brain lesion segmentation.

5.3 Multiple Sclerosis Data and Challenges

In this chapter, we use 78 MR volumes from 3 different local datasets for subjects with MS including ACTIMUS, SPM-MRC and MEDSCH datasets, with T1 and T2 FLAIR MRI sequences for each subject. ACTIMUS and SPM-MRC data were acquired on a 3T Achieva MR scanner (Philips Medical, Best, The Netherlands) using a 32 channel head coil. MEDSCH data were acquired on a 3T MR750 scanner (GE Healthcare, Milwaukee, USA) using a 32 channel head coil. Information about the number of samples and image dimensions are shown in Table 5.2.

Dataset	Samples	Minimum	Maximum
ACTIMUS	35	142, 170, 150	167, 224, 200
MEDSCH	23	146, 177, 153	175, 237, 200
SPMMRC	20	146, 171, 145	169, 212, 200

Table 5.2: Dataset details with the number of samples, minimum and maximum image dimensions, all with voxel dimensions of 1x1x1 millimetre.

5.3.1 preprocessing Methods

Before developing a method for MS lesion segmentation, it is important that input images are prepared for the segmentation task. The general

procedure is to extract the brain tissue from Skull and other remaining head structures and make sure all modalities and region of interest are properly aligned. In this section we compare two standard preprocessing approaches and compare the outcomes of these two methods:

Naive preprocessing Approach

The Naive Preprocessing (NP) procedure performed on the data is as follows: T2 FLAIR and T1 alignment using FMRIB’s Linear Image Registration Tool (FLIRT), a fully automated robust accurate tool for linear (affine) intra-and inter-modal brain image registration [Jenkinson and Smith, 2001, Jenkinson et al., 2002], Remove non-brain tissue using Brain Extraction Tool (BET) [Smith, 2002], Zero-mean normalisation.

Figure 5.3 shows the resulting T2 FLAIR and T1 volumes of target datasets using NP. Visual inspections reveal some issues in the NP procedure. For example, after brain extraction, for some of the subjects, the brain tissue was not extracted correctly. Figure 5.4 shows some of these inspected volume slices. The issues are marked by yellow arrows.

UK BioBank Pipeline preprocessing Approach

On the other hand, UK Biobank Preprocessing Pipeline (UKBP) [Alfaro-Almagro et al., 2018] for brain imaging is a fully automated processing pipeline primarily based on the FMRIB Software Library (FSL), a comprehensive library of analysis tools for functional, structural and diffusion MRI brain imaging data [Jenkinson et al., 2012]. UKBP is a robust pipeline that can process and integrate across many different functional and structural modalities. Therefore, this pipeline is used as the alternative preprocessing

procedure.

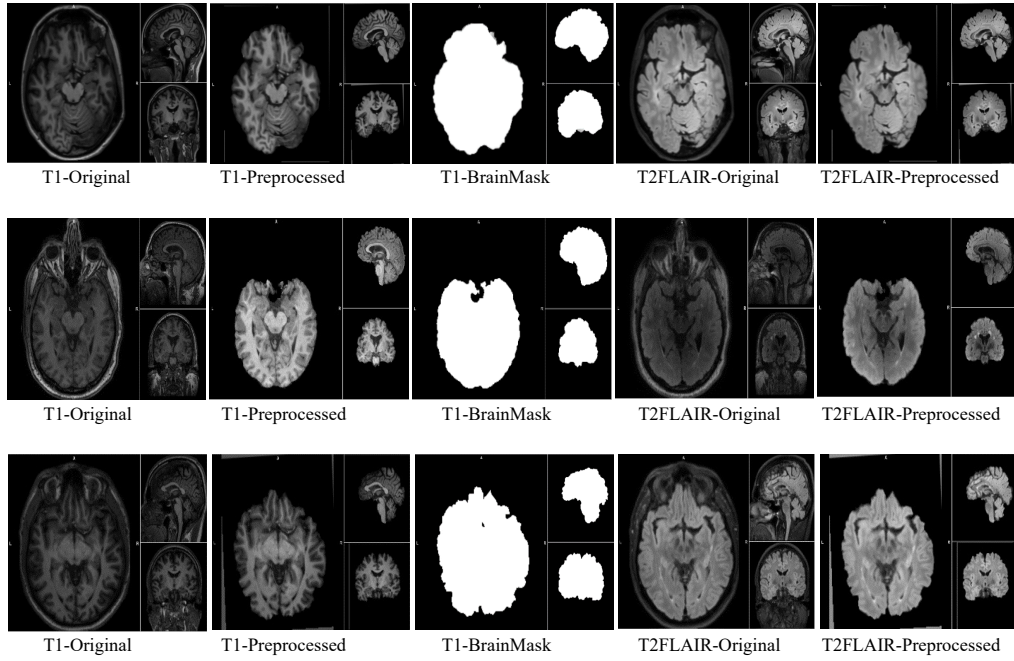


Figure 5.3: The original T1, T2 FLAIR and the resulting T1, T2 FLAIR and mask volumes by using the NP procedure. One subject per dataset are chosen for illustration, from top to bottom, subjects are selected from AC-TIMUS, MEDSCH and SPM-MRC datasets, respectively. For each image, the Axial, Coronal and Sagittal views are illustrated.

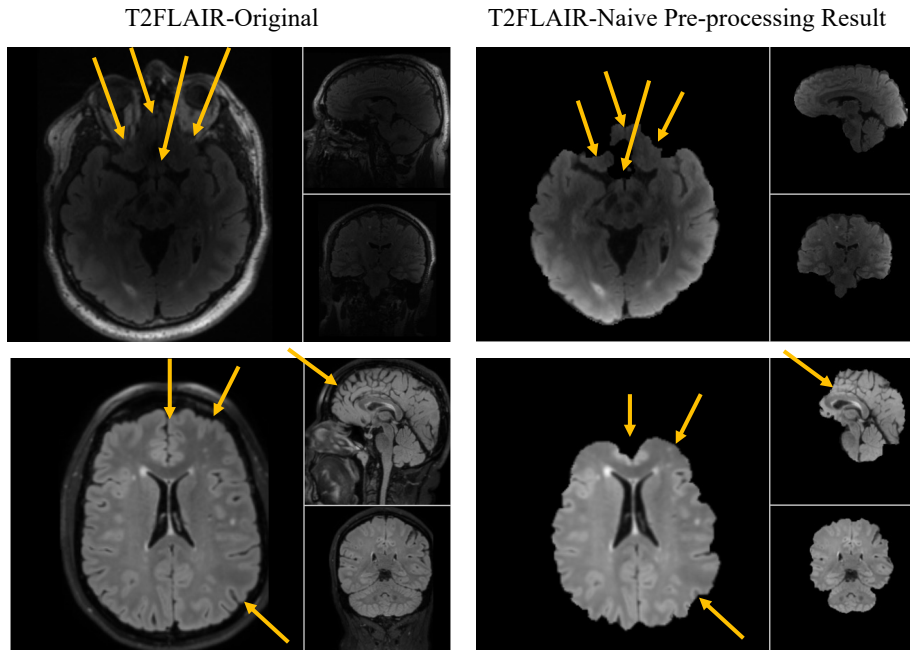


Figure 5.4: The NP results on some of the challenging T2 FLAIR images of MEDSCH and SPM-MRC datasets (from top to bottom). Parts of the brain tissue are not captured by NP as marked by the yellow arrows.

UKBP preprocessing performed on the data are as follows:

1. Identify the T1 brain start in z dimension using robustfov (image field of view truncation) and extract the region of interest,
2. Apply recursive BET on the T1 region of interest [Smith, 2002],
3. Reduce Field Of View (FOV) by aligning the T1 extracted brain to the standard space using the standard space reference image (MNI152T1),
4. Calculate affine transformation matrix for T1 to MNI standard space and non-linear transformation of T1 to MNI using the obtained matrix,
5. Generate the T1 brain mask,
6. T1 defacing by masking out extracranial voxels in the face and ear regions,
7. Linear registration of T2 FLAIR to T1 [Jenkinson and Smith, 2001, Jenkinson et al., 2002],
8. Generate the linear transformation matrix from T2 FLAIR to MNI standard space and T2 FLAIR defacing using the generated matrix,
9. Bias Field Correction on T1 and T2 FLAIR using FSL FAST [Zhang et al., 2001],
10. Ground Truth (GT) is aligned with the original T2 FLAIR. So, GT needs to be aligned with the new T2 FLAIR. To this end, we added another step to the standard UKBP pipeline. A linear registration is performed from GT to T2 FLAIR by applying the transformation matrix generated from T2FLAIROriginal to T2FLAIRFinal and using the reference volume T1 Extracted Brain,

11. Zero-mean normalisation is done after the UKBP preprocessing.

Figure 5.5 and Figure 5.6 show the resulting T2 FLAIR and T1 volumes of target datasets using UKBP preprocessing, and some sample slices of the aligned ground truth, respectively. It can be seen from Figure 5.6 that the preprocessed ground truth is fully aligned with the lesions in the T2 FLAIR slices.

Figure 5.7 shows a comparative illustration of the two preprocessing procedures. A comparison over the visual inspections of the UKBP preprocessing shows that UKBP in most cases could successfully extract the tissues of the brain that were not extracted by the Naive preprocessing.

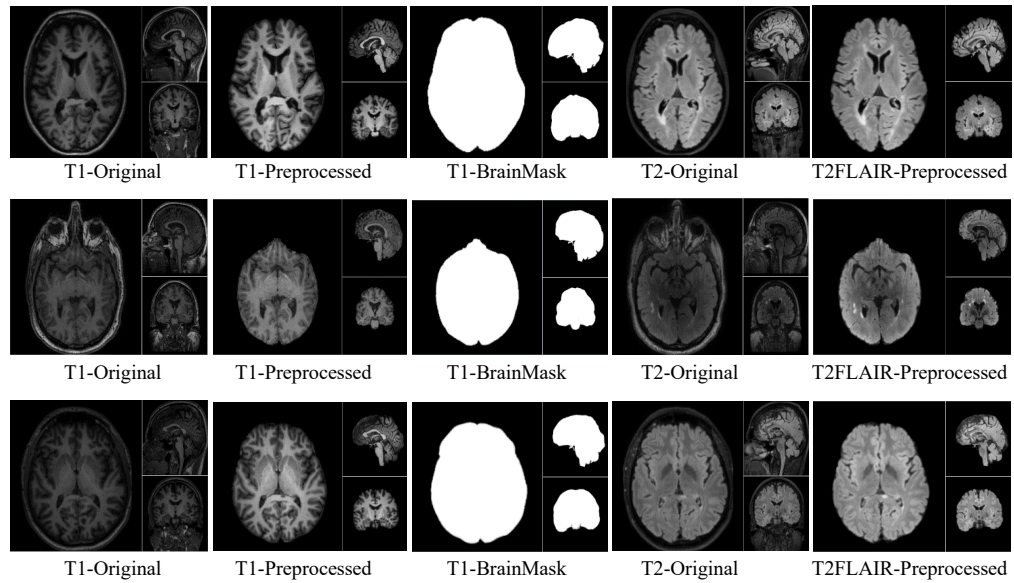


Figure 5.5: The original T1, T2 FLAIR and the resulting T1, T2 FLAIR and mask volumes by using the UKBP procedure. One subject per dataset are chosen for illustration, from top to bottom, subjects are selected from ACTIMUS, MEDSCH and SPMMRC datasets, respectively. For each image, the Axial, Coronal and Sagittal views are illustrated.

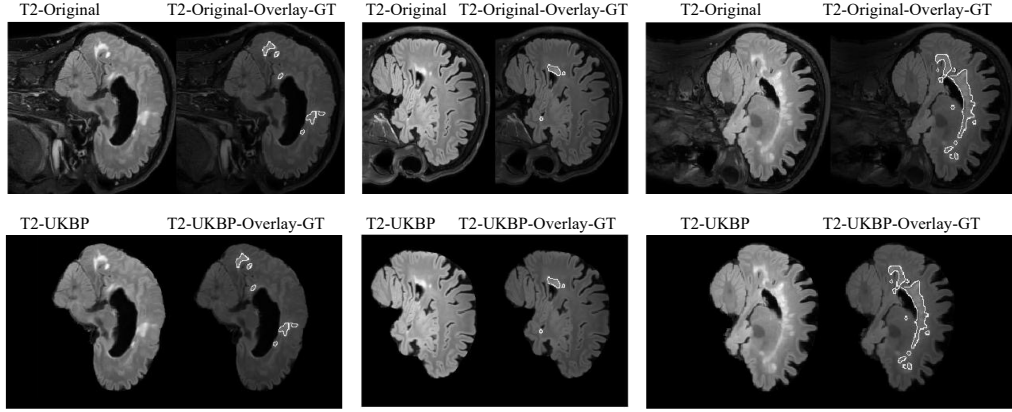


Figure 5.6: Ground truth alignment; the top row shows slices of three samples from their original volumes with and without contoured ground truths. The bottom row shows the respective results from The Modified UKBP preprocessing with and without contoured ground truths.

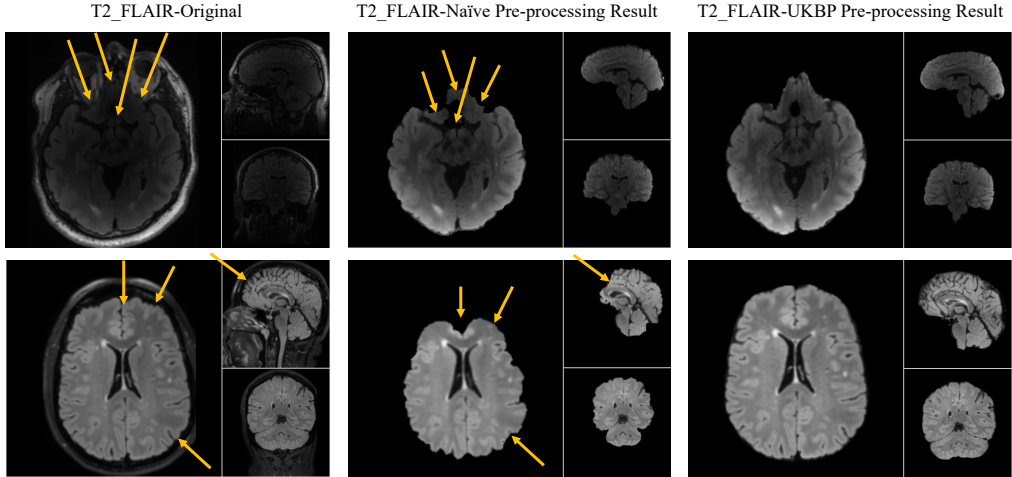


Figure 5.7: The comparison between NP and UKBP; each row shows the original, NP and UKBP T2 FLAIR slices of a subject, respectively.

5.4 Experiments

As discussed earlier, our aim in this chapter is to employ our current knowledge and ideas to MS medical image segmentation. We explored the approaches to preparing MS images for segmentation. We aim to conduct experiments with our three in-house MS datasets. Our experiments involves comparing the effect of the two preprocessing approaches (NP and

UKBP) on the segmentation performance. Then, using the best performing preprocessing method, our next objective is to assess the effectiveness of our tailored loss functions (UCF, ELF, UCF+AL and ELF+AL introduced in Chapter 3) in the context of Fully supervised MS brain lesion segmentation.

Therefore, methods selected for all experiments in this chapter are as follows:

Multi-Scale 3D Convolutional Neural Network (DeepMedic)

DeepMedic is a 3D Convolutional Neural Network tailored for segmentation tasks, specifically targeting brain lesion segmentation [Kamnitsas et al., 2015]. The network consists of 11 layers, with a double-pathway structure. This dual-pathway approach allows for the simultaneous processing of 3D patches at multiple scales, to enhance the model’s ability to discern fine details and contextual information essential for accurate segmentation. The reason for selecting DeepMedic as one of the reference models in this chapter is its ability to be effectively trained on a small dataset. This is particularly important in medical imaging, where large annotated datasets are often scarce. Furthermore, their study shows that the double-pathway architecture facilitates the analysis of 3D patches at multiple scales by processing both the original and a subsampled version of an image to provide a broader context for each voxel, enhancing the model’s segmentation accuracy.

Multi modality Unet (MUnet)

The second reference model selected, for evaluating our proposed tailored loss functions for fully supervised segmentation in this chapter, we use a MUnet with a similar structure as explained in Chapter 3 with the network accepting two modalities (T1 and T2 FLAIR) as input. MUnet is designed to accept 2D slices of the 3D MRI images as input.

The reader is referred to the Appendix B for the model training details and parameter settings.

The Training, Validation, and Test sets chosen for the experiments we name 'Multi-Site Mixture' and 'Single-Site Mixture'. Depending on the subjects used to train the reference models, we use these two mixtures throughout this chapter. The details of these two mixtures are shown in Table 5.3. The reason for choosing these sets will be explained in the corresponding experimental section.

Before presenting the results of our experiments, it is important to note that, due to the use of an in-house dataset, direct comparisons with results from other studies are limited. These other studies often employed various public and in-house datasets with different protocols and scanners. Based on a recent study [Kamraoui et al. [2022]] that compared well-known MS lesion segmentation methods—including *nicMSlesion* [Valverde et al., 2019], *DeepMedic* [Kamnitsas et al., 2016], *2.5D Tiramisu* [Zhang et al., 2019a], and *DLB* [Kamraoui et al., 2022]—a range of Dice scores achieved by these state-of-the-art models using different training datasets is provided in Table 5.4.

When these models were trained on one dataset and tested on another, Dice scores ranged from 0.204 to 0.746. Statistical tests revealed no significant

Multi-Site Mixture			
Dataset	Training Subjects	Validation Subjects	Test Subjects
ACTIMUS	26	2	7
SPMMRC	14	2	4
MEDSCH	16	2	5

Single-Site Mixture			
Dataset	Training Subjects	Validation Subjects	Test Subjects
ACTIMUS	30	5	-
MEDSCH	-	-	23
SPMMRC	-	-	20

Table 5.3: The Training, Validation, and Test sets chosen from the three local datasets (ACTIMUS, SPMMRC and MEDSCH) for the experiments. As can be seen, the Single-Site mixture is so-called as it uses data from only a single site in training and validation.

Scenario	Dice Range	Statistical Test Result
Train on Dataset A / Test on Dataset B	0.204 - 0.746	No significant difference
Train on Dataset A / Test on Dataset A	0.639 - 0.672	No significant difference

Table 5.4: Summary of Dice score ranges for MS lesion segmentation models based on training and testing scenarios reported in the literature [Kamraoui et al., 2022]. The table shows the range of Dice scores achieved when models were trained on one dataset and tested on another (cross-dataset) versus when trained and tested on the same dataset. No statistically significant differences were observed in either scenario.

differences among the models in this cross-dataset scenario. Conversely, when both the training and test sets were drawn from the same dataset, the Dice scores for these models ranged from 0.639 to 0.672, again with no statistically significant differences observed.

This variation in Dice scores highlights the sensitivity of these models to

domain shift issue. Depending on the dataset used for training, the models may exhibit varying degrees of robustness to differences in data distribution, scanner types, and imaging protocols. In challenging cases, where domain shift is more pronounced, Dice scores can be quite low, reflecting the inherent difficulties in generalising across different datasets. The differences in scanners and protocols further complicate the segmentation task, as these factors can impact the models' performance, making it harder to achieve consistent results across diverse datasets.

5.4.1 Preprocessing Pipeline Comparative Experiments

To understand the differences between the two preprocessing methods discussed in Section 5.3.1, a comprehensive set of experiments was designed on our local MS datasets. Reference models for these experiments are DeepMedic and MUnet with a cross entropy loss function. The parameter settings are explained in appendix B. For this set of experiments, we used the Multi-Site Mixture shown in Table 5.3. The subjects are chosen from a mixture of all three local sets for train, validation and test in order to use more data for training the model and focus on evaluating the performance of the preprocessing methods. We trained the models on data preprocessed using NP and UKBP preprocessing methods.

Results

Table 5.5 shows the comparison of the trained models in terms of dice coefficient performance measure in our designed experiments. Both the DeepMedic model and MUnet show higher performance in Dice Coefficient when utilising UKBP preprocessing. These results support the visual re-

sults presented in the previous section and, for this reason, we will use UKBP as the preprocessing approach going forward.

Model	Preprocessing Method	Dice Coefficient
DeepMedic	NP	0.597 $\pm$ 0.045
DeepMedic	UKBP	0.70$\pm$0.029
MUnet	NP	0.673 $\pm$ 0.034
MUnet	UKBP	0.69$\pm$0.028

Table 5.5: MS lesion segmentation performance (on 16 Test Subjects) comparison of preprocessing methods NP and UKBP for the baseline models DeepMedic and MUnet using Multi-Site Mixture data.

5.4.2 MS Lesion Segmentation Tailored Loss Function Comparative Experiments

In this section, fully-supervised MS lesion segmentation experiments are designed. The reference models are DeepMedic and MUnet, as described above. The loss functions proposed in Chapter 3 with the MUnet reference model are used in this section for the segmentation of MS lesions. The performance of the models is compared with DeepMedic and MUnet with cross-entropy loss as baseline. The models are trained on the Multi-site mixture set, where a combination of all three local datasets (ACTIMUS, MEDSCH and SPMRC) are involved in training, validation and testing. The Multi-site mixture set was selected for this experiment to ensure that the model is exposed to images from all three in-house datasets during both training and testing phases, thus removing the influence of image acquisition variability in this specific experiment.

Results

Table 5.6 shows the experiments’ results. Based on the calculation of the Dice Coefficient, MUnet+UCF performs best, but closely followed by DeepMedic. The baseline MUnet model performed slightly below DeepMedic, which aligns with expectations given DeepMedic’s more complex architecture. The fact that MUnet + UCF closely matched and even slightly surpassed DeepMedic underscores the importance of loss function design in segmentation tasks. The effectiveness of UCF, ELF, and AL regularisation term could vary depending on the characteristics of the data and the dataset used for training. For datasets with high noise or variability, AL is important as it helps the model become more robust against inherent data imperfections. UCF, on the other hand, is more beneficial when dealing with high epistemic uncertainty, where the model lacks knowledge or exposure to varied examples. In contrast, ELF focuses on iterative error correction and can enhance performance by learning from previous mistakes. The interplay between these loss functions is tailored to the specific data context and in some scenarios the effect of AL could cause too much regularisation the results in underfitting.

Model	Dice Coefficient
DeepMedic	0.70±0.029
MUnet (Baseline)	0.69±0.028
MUnet + UCF	0.702±0.021
MUnet + UCF + AL	0.678±0.029
MUnet + ELF	0.675±0.033
MUnet + ELF + AL	0.683±0.026

Table 5.6: MS lesion segmentation performance comparison on tailored loss functions and the baseline models DeepMedic and MUnet using UKBP pipeline with Multi-Site mixture data.

Figure 5.8 shows the prediction results in three samples selected from the test set. Each row illustrates a selected slice of the prediction results for each model overlayed on the corresponding T2FLAIR image. The visual inspection although very challenging to interpret due to the distribution of the lesions throughout the brain with most of them being small lesion, shows that the models' lesion segmentation performance are very close. The images for inspection are chosen randomly.

5.4.3 Understanding the Impact of Acquisition Site and Scanner on MS Lesion Segmentation Model Performance

The purpose of this set of experiments is to evaluate the robustness of segmentation methods when the test data are collected from different sites and/or scanners than the images used during the training phase. For these experiments, the reference models are DeepMedic and MUnet. The loss functions proposed in Chapter 3 with the MUnet reference model are used in this section for the segmentation of MS lesions. The performance of the models is compared with DeepMedic and MUnet with a cross-entropy loss as baseline. The models are trained on the Single-site mixture set, where only the ACTIMUS dataset is involved in training and validation, and SPMARC and MEDSCH are used as a test set (details shown in Table 5.3).

Results

The segmentation results with Dice Coefficient are shown in Table 5.7. Three Dice Coefficient Values are reported for each model, one total aver-

age dice, and one for each of the test sets, SPMMRC and MEDSCH calculated separately. The MUnet+UCF+AL model achieved the highest total average Dice and highest Dice on the MEDSCH set. The MUnet+UCF achieved the highest Dice on the SPMMRC test set.

Between MEDSCH and SPMMRC sets, the SPMMRC set has a higher performance in terms of Dice. This is due to the fact that although the Acquisition site of training set (ACTIMUS) is different from SPMMRC, the scanner used for both of these sets are similar (scanner details are mentioned in section 5.3). On the other hand, the performance of all models is lower on MEDSCH, due to the fact that MEDSCH images were acquired from a different scanner than the images in the training set. Figure 5.9 shows the segmentation results in three samples selected from the test set. Each row illustrates a selected slice of the prediction results for each model overlayed on the corresponding T2FLAIR image. The visual inspection of the images in comparing the best proposed models and DeepMedic confirms better segmentation for UCF+AL specifically for MEDSCH testset. To evaluate the statistical significance of the values for each performance measure, we applied the non-parametric Wilcoxon signed rank statistical test [Gibbons and Chakraborti, 2011].

While incorporating the AL regularisation term and the attention-based UCF loss did show some potential for improving Dice coefficients and enhancing model generalisation, the observed improvements were not statistically significant. This suggests that although AL and UCF can offer benefits, they may not fully address the domain shift issue on their own. Therefore, while these loss functions indicate potential avenues for improving model robustness and generalisability, more comprehensive analysis and further research are needed to better understand and address the challenges associated with domain shift.

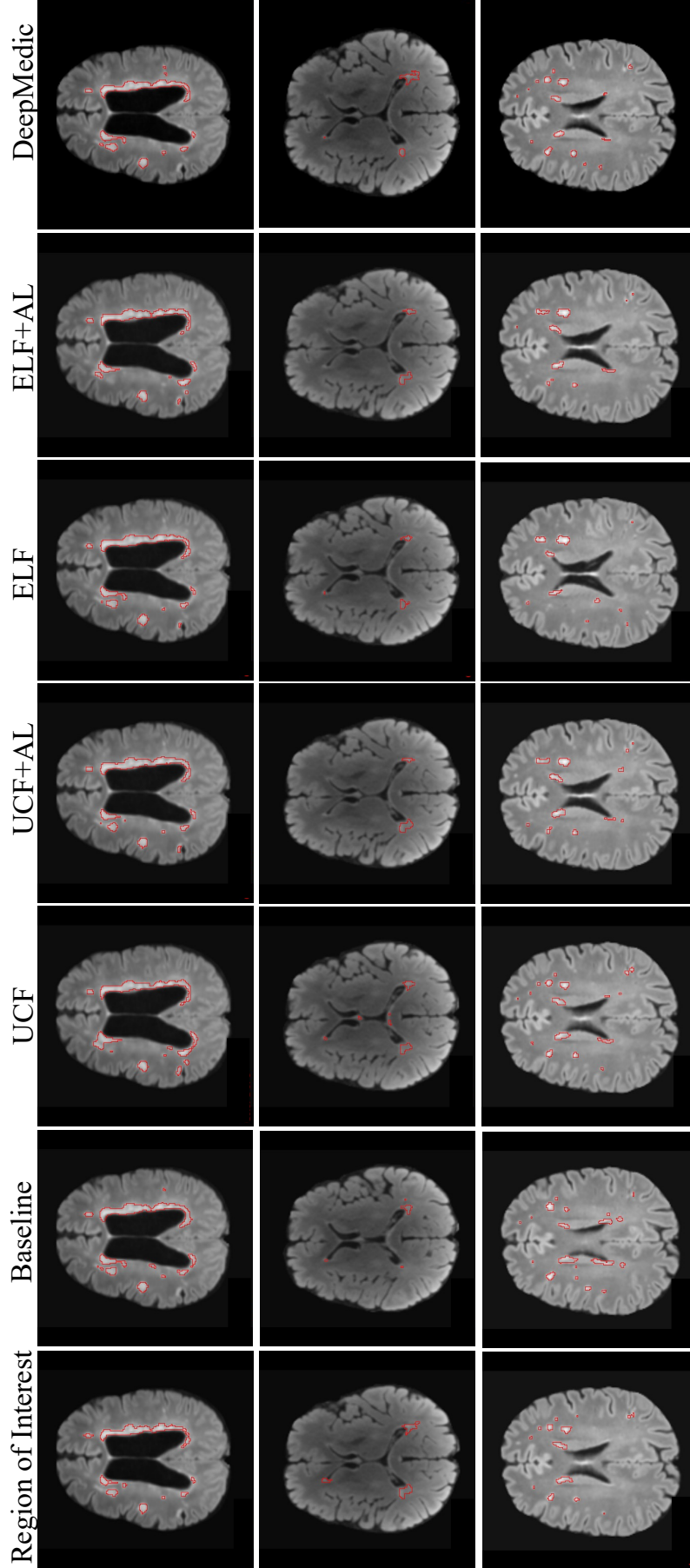


Figure 5.8: Comparative visualisation of predictive performance across three distinct test samples: MUnet employing baseline cross-entropy loss (Baseline), MUnet with the proposed tailored loss functions, and DeepMedic, all predictions are overlaid on T2 FLAIR images, displayed in axial view. Each row corresponds to the results from a single sample, showcasing the efficacy of each approach.

Model	Total Dice	MEDSCH Dice	SPMMRC Dice
DeepMedic	0.505 ± 0.059	0.377 ± 0.053	0.652 ± 0.025
MUnet (Baseline)	0.436 ± 0.069	0.279 ± 0.055	0.617 ± 0.024
MUnet+UCF	0.470 ± 0.068	0.3 ± 0.050	0.666 ± 0.016
MUnet+UCF+AL	0.511 ± 0.049	0.398 ± 0.05	0.641 ± 0.018
MUnet+ELF	0.453 ± 0.07	0.312 ± 0.066	0.615 ± 0.028
MUnet+ELF+AL	0.492 ± 0.063	0.356 ± 0.061	0.648 ± 0.02

Table 5.7: MS lesion segmentation performance comparison on learning from Single-Site Mixture and evaluating on Images acquired from different site/scanner; Preprocessing using UKBP pipeline for the baseline models DeepMedic and MUnet and tailored loss functions.

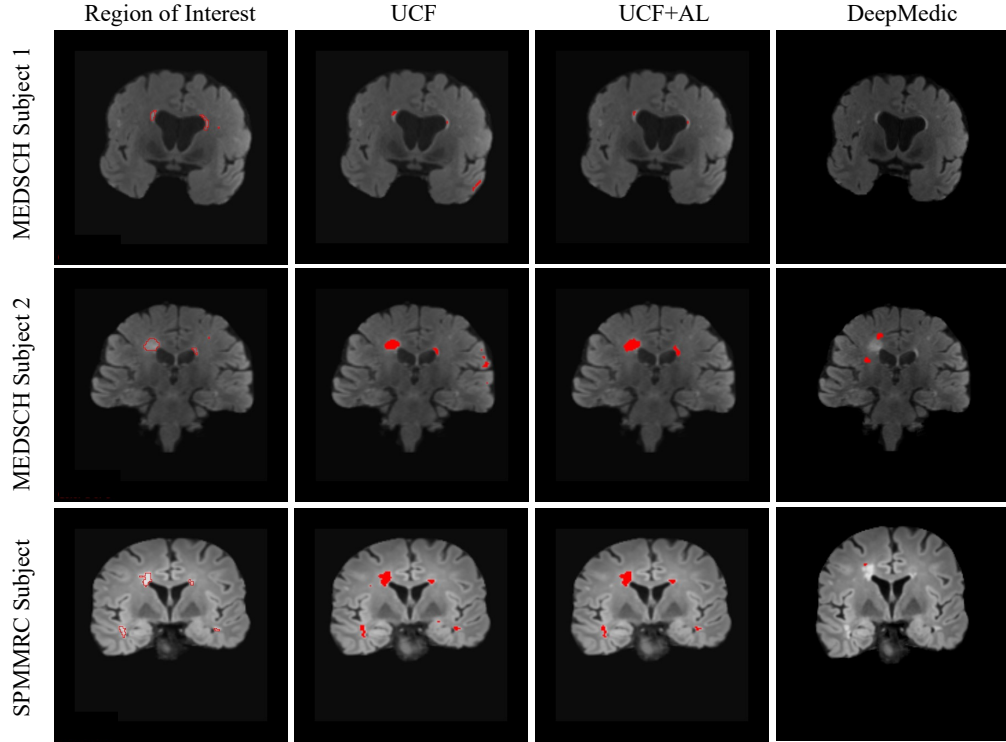


Figure 5.9: Comparative visualisation of predictive performance across three distinct test samples: MUnet with the best proposed tailored loss functions (UCF and UCF+AL), and DeepMedic, all predictions are overlaid on T2 FLAIR images, displayed in coronal view. Each row corresponds to the results from a single sample, showcasing the efficacy of each approach.

5.5 Discussion

As discussed in Chapter 3, medical image segmentation is a challenging task due to the complexity of the data and the complex nature of uncertainty in the images. We discussed the factors that can affect the segmentation performance. One of these influencing factors is inconsistencies in acquisition methods and the role of tailored loss functions, especially in handling out-of-distribution data.

In this chapter, we demonstrated that in particularly complex data such as MS MRI images, even choosing the right pipeline for preprocessing the images and extracting the brain tissue from the whole image can have a notable impact on the outcome of the subsequent segmentation method. This effect was more obvious in the DeepMedic reference model. The DeepMedic model was shown to be more sensitive to the preprocessing method.

In comparing the results of the proposed tailored loss functions with an MUnet as baseline, UCF loss and DeepMedic performed similarly despite the fact that for MUnet+UCF method, the images were downsized due to computational reasons and the model was only trained on 2D slices of the MRI images.

Deep Medical Image segmentation methods have shown limited application in MRI for tasks like MS lesion segmentation [Kamraoui et al., 2022], primarily due to their inability to generalise well to new, unseen data from medical sites not included in their training datasets, as shown in the experiments with Single-Site Mixture set. This challenge stems from the inherent diversity and variability in real-world data; the data are not curated and collected in a homogenous way. Factors such as optimisation choices, the heterogeneity of training data, and the features extracted can

affect a model’s generalisation capabilities [Kamraoui et al., 2022]. In the experiments of this chapter, the performance of the models was lower on the images acquired from a different scanner than when some images from the site/scanner are included in the training. However, using the UCF+AL loss is shown to be more effective in the robustness of the model when segmenting these unseen data. As discussed in chapter 3, variations arising from the use of various scanners for data collection may lead to inherent uncertainties in the data. The presence of such variations, if not encountered by the model during training, could result in the introduction of model uncertainty. By incorporating an attention mechanism driven by uncertainty during training, the model with UCF+AL loss could effectively manage these variations.

5.6 Conclusion

In diagnosing MS, an inflammatory demyelinating disease of the central nervous system, MRI plays a pivotal role. MRI’s sensitivity in detecting MS lesions, characterised by disruptions in the brain and spinal cord, is crucial for confirming the disease’s presence, understanding its progression, and assessing treatment efficacy. T2 FLAIR imaging, recommended for identifying various MS lesion types, and T1-weighted imaging, through 3D MPRAGE, are instrumental in this process. The identification of the CVS within lesions has emerged as a promising diagnostic tool, given its prevalence in MS lesions compared to other conditions. One of the important steps towards CVS detection is accurate lesion segmentation. Despite advances, the efforts for accurate MS lesion segmentation and the reduction of misdiagnosis remains a critical challenge, underscoring the need for refined imaging biomarkers and the importance of robust lesion segmenta-

tion methodologies in enhancing diagnostic precision and patient care in MS.

In this chapter, we evaluated two preprocessing approaches, NP and UKBP, to extract brain tissue from the remaining head structure. We found UKBP a more robust approach by visual inspection of the preprocessed images and by evaluating the segmentation performance of two reference models on preprocessed NP and UKBP images.

Furthermore, we assessed the performance of our tailored loss functions for fully supervised segmentation proposed in Chapter 3. Our experiments on the Multi-Site Mixture dataset show that MUnet+UCF performs best, but closely followed by DeepMedic.

In a further set of particularly challenging experiments, using the Single-Site Mixture dataset, we demonstrated that MUnet+UCF+AL performs better than other methods, specifically when tested on the MEDSCH dataset where the acquisition scanner for the images were different from the images included in the training set. Our experiments show that utilising loss functions that feedback the uncertainty of the model and use the inherent data uncertainty as regularisation term can improve the performance of segmentation when the model is tested on unseen data from various sites and scanners. This underscores the significance of creating a custom loss function to enhance the robustness of the model, alongside the architectural design of the Deep MS lesion segmentation model.

Chapter 6

Conclusions

In this chapter the key contributions and conclusions of this thesis are discussed. In this thesis, two fundamental challenges in deep medical image segmentation, the *output reliability and interpretability problem*, and the *strong annotation scarcity problem* were explored. To address these challenges, this thesis proposed to build tailored loss functions for fully and weakly supervised deep medical image segmentation. The following solutions were explored in this thesis:

Uncertainty-Aware Loss Function for Fully Supervised Medical Image Segmentation: This proposed solution, focused on the interpretability of uncertainty estimation in BNNs specifically for medical image segmentation. It explored how various factors, including lesion characteristics and imaging inconsistencies, impact the segmentation process. It introduced unique loss functions designed to incorporate uncertainty estimation directly into the training of fully supervised DCNN models. These include the UCF, ELF, UCF+AL and ELF+AL. These tailored loss functions addressed model and data uncertainties via different mechanisms,

enhancing a model’s interpretability and accuracy.

Extensive experiments were conducted on three benchmark medical imaging datasets, each chosen for its distinct challenges and characteristics, to evaluate the proposed loss functions and their impact on uncertainty interpretation. The analysis across these datasets revealed that the ELF model, particularly when augmented with Aleatoric Loss (ELF+AL), achieved superior segmentation accuracy, as evidenced by higher Dice scores. This finding was consistent across two of the benchmarks, ISIC2018 and BraTS, with visual inspections confirming the quantitative results. The ELF+AL model, in particular, showed improvement in uncertainty heatmaps, highlighting its effectiveness in managing uncertainties.

The findings of this solution underscore the importance of tailored uncertainty management in medical diagnostics, where accurate lesion segmentation and uncertainty interpretation are critical for clinical decision-making. The solution demonstrates that the choice of loss function significantly affects the system’s ability to handle unseen data and adapt to out-of-distribution examples, thereby enhancing the reliability of automated diagnostic systems.

Uncertainty-Guided and Spatially-Constrained Weakly Supervised Deep Medical Image Segmentation: This proposed solution delves into the development and assessment of a novel approach, for skin lesion segmentation in medical images, specifically utilising the ISIC2018 dataset for the experiments. By employing bounding box annotations as a guide, the model iteratively improves its predictions towards the ground truth, facilitated by a custom loss function that integrates weighted cross-entropy with CRF terms. This novel loss function employs both epistemic uncer-

tainty and spatial data relationships, enhancing the model’s confidence and performance in identifying lesion areas.

The epistemic uncertainty heatmaps serve as a critical guide, revealing the model’s evolving confidence across training epochs and highlighting persistently challenging areas for segmentation. Such insights underscore the method’s dynamic adaptation and the potential for further refinement, especially in accurately delineating complex lesion boundaries. The method demonstrates varying effectiveness depending on the lesion’s contrast against the background, with the CRF playing a significant role in reinforcing segmentation accuracy, albeit with some limitations in cases of subtle intensity variations.

The proposed method represents a performance improvement in weakly supervised skin lesion segmentation, offering an innovative solution that combines epistemic uncertainty estimation and spatial constraints. However, it is important to note that challenges remain, particularly in handling lesions with minimal intensity differences from their background. These insights pave the way for future improvements, especially in optimising the CRF component to better manage such subtle variations, ensuring more reliable and precise segmentation outcomes in medical diagnostics.

MS Lesion Segmentation Case Study: A significant focus of this thesis is a case study on MS lesion segmentation, aiming to delve into the complexities and challenges inherent in segmenting MS lesions from MRI images. This study specifically addresses the difficulties posed by the challenging nature of MS presentation, and the variability in image acquisition methods, underscoring the importance of designing tailored loss functions to effectively manage these challenges, especially with data that

diverges from what model has encountered.

This case study explored the critical role of preprocessing techniques in enhancing the accuracy of segmenting brain lesions, noting that the selection of the preprocessing pipeline can significantly influence the results. It further investigates the performance of various loss functions in improving segmentation accuracy, particularly for data acquired from different scanners and locations, which often present unseen challenges. Despite computational constraints that necessitated modifications such as image downsizing and training on 2D slices, the proposed tailored loss functions (UCF and UCF+AL) demonstrated to be effective, in comparison to approaches that focus on modifying the DCNN architecture itself (DeepMedic [Kamnitsas et al., 2016]).

This case study highlights the sensitivity of the DeepMedic reference model to preprocessing methods. This sensitivity underscores the need for careful choice of preprocessing pipelines to optimise segmentation outcomes. Moreover, the study reveals that tailored loss functions, particularly UCF+AL, enhance the model’s robustness in segmenting unseen data, a crucial factor given the diversity and variability of real-world medical data and the challenge of generalising across different medical imaging sites and scanners.

Summary In summary, this thesis presented novel solutions to the critical challenges of output interpretability, along with accurate annotation scarcity in medical image segmentation on both challenging benchmark datasets and in-house datasets. By developing tailored loss functions, this work, particularly contributed to the fully and weakly supervised segmentation. The introduction of uncertainty-aware loss functions and spatial constraints, shown improvement in segmentation performance and inter-

pretability of segmentation results.

Future Work As a future direction, methods that use global constraints on the model such as size of the lesion by simply putting constraints on softmax, in a form of regularisation in loss function should be explored due to their success in improving the performance of segmentation models [Kervadec et al., 2019, Jia et al., 2017, Bateson et al., 2019, Pathak et al., 2015].

Another future direction, based on our preliminary studies, is the use of topological Data Analysis (TDA) tools to translate shape priors to loss function. Topology, a branch of mathematics, focuses on the study of shapes and their intrinsic properties that remain unchanged under continuous transformations, such as stretching or bending. Topological concepts provide powerful tools for understanding abstract and complex spatial relationships that extend beyond simple geometric measurements Carlsson [2009, 2020]. The increasing complexity of datasets in various scientific domains has led to the development of TDA, an innovative approach that incorporates topological concepts for the analysis of large and intricate data sets.

TDA models data using graphs or networks instead of traditional linear algebraic or cluster analysis methods, capturing the shape of the data in a combinatorial manner. This approach leverages mathematical methods, such as homology, to measure shape and gain insights into the overall structure of the data set Carlsson [2009, 2020]. Approaches such as Semi-supervised learning for segmentation often incorporate prior/domain knowledge, such as anatomical information about the objects of interest, as strong regularisation. This prior knowledge, including location,

shape, anatomy, and context, is crucial for manual annotation, particularly when boundaries are unclear or image contrast is low. In deep network-based semantic segmentation, training typically relies on local or pixel-wise loss functions, which might not adequately capture underlying anatomical structures and global dependencies. Anatomical-prior-guided methods assume the solution space can be expressed as a prior distribution, encouraging the network to produce more anatomically plausible segmentation Peng and Wang [2021].

Bibliography

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Abolfazl Abdollahi, Biswajeet Pradhan, and Abdullah Alamri. Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access*, 8:179424–179436, 2020.

Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

Ayelet Akselrod-Ballin, Meirav Galun, John Moshe Gomori, Massimo Filippi, Paola Valsasina, Ronen Basri, and Achi Brandt. Automatic segmentation and classification of multiple sclerosis in multichannel mri. *IEEE transactions on biomedical engineering*, 56(10):2461–2469, 2009.

- O Aktas, MP Wattjes, M Stangel, and H-P Hartung. Diagnosis of multiple sclerosis: revision of the mcdonald criteria 2017. *Der Nervenarzt*, 89: 1344–1354, 2018.
- Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.
- Devansh Arpit, Víctor Campos, and Yoshua Bengio. How to initialize your network? robust initialization for weightnorm & resnets. *Advances in Neural Information Processing Systems*, 32, 2019.
- S Bakas, H Akbari, A Sotiras, M Bilello, M Rozycki, J Kirby, J Freymann, K Farahani, and C Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection (2017). DOI: <https://doi.org/10.7937/K>, 9.
- Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017a.
- Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017b.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Constrained domain adaptation for segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 326–334. Springer, 2019.

Pierre-Louis Bazin and Dzung L Pham. Statistical and topological atlas based brain image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 94–101. Springer, 2007.

Pierre-Louis Bazin and Dzung L Pham. Homeomorphic brain image segmentation with topological and statistical atlases. *Medical image analysis*, 12(5):616–625, 2008.

Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

Cristiane Benvenuto-Andrade, Stephen W Dusza, Anna Liza C Agero, Alon Scope, Milind Rajadhyaksha, Allan C Halpern, and Ashfaq A Marghoob. Differences between polarized light dermoscopy and immersion contact dermoscopy for the evaluation of skin lesions. *Archives of dermatology*, 143(3):329–338, 2007.

- Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.
- John-Melle Bokhorst, Hans Pinckaers, Peter van Zwam, Iris Nagtegaal, Jeroen van der Laak, and Francesco Ciompi. Learning from sparsely annotated data for semantic segmentation in histopathology images. 2018.
- Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.
- Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001.
- John Bradshaw, Alexander G de G Matthews, and Zoubin Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv preprint arXiv:1707.02476*, 2017.
- Tom Brosch, Roger Tam, and Alzheimer’s Disease Neuroimaging Initiative. Manifold learning of brain mris by deep learning. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22–26, 2013, Proceedings, Part II 16*, pages 633–640. Springer, 2013.
- Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Troubsee, and Roger Tam. Deep 3d convolutional encoder networks

- with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.
- Wallace J Brownlee, Todd A Hardy, Franz Fazekas, and David H Miller. Diagnosis of multiple sclerosis: progress and challenges. *The Lancet*, 389(10076):1336–1346, 2017.
- Graeme M Bydder and Ian R Young. Mr imaging: clinical use of the inversion recovery sequence. *J Comput Assist Tomogr*, 9(4):659–675, 1985.
- Jinzheng Cai, Youbao Tang, Le Lu, Adam P Harrison, Ke Yan, Jing Xiao, Lin Yang, and Ronald M Summers. Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3d mask generation from 2d recist. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 396–404. Springer, 2018.
- Yigit B Can, Krishna Chaitanya, Basil Mustafa, Lisa M Koch, Ender Konukoglu, and Christian F Baumgartner. Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 236–244. Springer, 2018.
- Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.
- Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

- Gunnar Carlsson. Topological methods for data modelling. *Nature Reviews Physics*, 2(12):697–708, 2020.
- Hang Chang, Qing Yang, and Bahram Parvin. A bayesian approach for image segmentation with shape priors. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- Yian-Leng Chang and Xiaobo Li. Adaptive image region-growing. *IEEE transactions on image processing*, 3(6):868–872, 1994.
- Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020a.
- Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- Tzu-Hsiu Chen, Shu-Hui Wang, Lin-Hui Su, Yu-Ling Hsu, Tsung-Hua Tsai, Ya-Jing Hsu, and Ying-Jui Chang. Comparison of visual effects of immersion fluids for dermoscopic examination of acral volar melanocytic lesions. *Dermatologica Sinica*, 32(2):69–74, 2014.
- Zhe Chen, Zhao Chen, Jingxin Liu, Qiang Zheng, Yuang Zhu, Yanfei Zuo, Zhaoyu Wang, Xiaosong Guan, Yue Wang, and Yuan Li. Weakly

- supervised histopathology image segmentation with sparse point annotations. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1673–1685, 2020b.
- Xi Cheng, Li Zhang, and Yefeng Zheng. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252, 2018.
- Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.
- Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.
- Guy Barrett Coleman and Harry C Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.
- Pierrick Coupé, Pierre Yger, Sylvain Prima, Pierre Hellier, Charles Kervrann, and Christian Barillot. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *IEEE transactions on medical imaging*, 27(4):425–441, 2008.
- Antonios Danelakis, Theoharis Theoharis, and Dimitrios A Verganelakis. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 70:83–100, 2018.

- Benoit M Dawant. Non-rigid registration of medical images: purpose and methods, a short survey. In *Proceedings IEEE International Symposium on Biomedical Imaging*, pages 465–468. IEEE, 2002.
- Bob D De Vos, Jelmer M Wolterink, Pim A De Jong, Max A Viergever, and Ivana Išgum. 2d image classification for 3d anatomy localization: employing deep convolutional neural networks. In *Medical imaging 2016: Image processing*, volume 9784, pages 517–523. SPIE, 2016.
- Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.
- Ivana Despotović, Bart Goossens, Wilfried Philips, et al. Mri segmentation of the human brain: challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015, 2015.
- Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018.
- Atam P Dhawan. *Medical image analysis*. John Wiley & Sons, 2011.
- Idanis Diaz, Pierre Boulanger, Russell Greiner, and Albert Murtha. A critical review of the effects of de-noising algorithms on mri brain tumor segmentation. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3934–3937. IEEE, 2011.
- Jordan D Dworkin, Pascal Sati, Andrew Solomon, Dzung L Pham, Richard Watts, Melissa L Martin, Daniel Ontaneda, Matthew K Schindler, Daniel S Reich, and Russell T Shinohara. Automated integration of

multimodal mri for the probabilistic detection of the central vein sign in white matter lesions. *American Journal of Neuroradiology*, 39(10):1806–1813, 2018.

Simon F Eskildsen, Pierrick Coupé, Vladimir Fonov, José V Manjón, Kelvin K Leung, Nicolas Guizard, Shafik N Wassef, Lasse Riis Østergaard, D Louis Collins, Alzheimer’s Disease Neuroimaging Initiative, et al. Beast: brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362–2373, 2012.

Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Xue Feng, Nicholas J Tustison, Sohil H Patel, and Craig H Meyer. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in computational neuroscience*, 14:25, 2020.

Massimo Filippi, Paolo Preziosa, Brenda L Banwell, Frederik Barkhof, Olga Ciccarelli, Nicola De Stefano, Jeroen JG Geurts, Friedemann Paul, Daniel S Reich, Ahmed T Toosy, et al. Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain*, 142(7):1858–1875, 2019.

Maggie A Flower. *Webb’s physics of medical imaging*. CRC press, 2012.

Yarin Gal. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

- Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589, 2023.
- Mina Ghaffari, Arcot Sowmya, and Ruth Oliver. Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the brats 2012–2018 challenges. *IEEE reviews in biomedical engineering*, 13:156–168, 2019.
- Mohsen Ghafoorian and Bram Platel. Convolutional neural networks for ms lesion segmentation, method description of diag team. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.
- Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric statistical inference*. Springer, 2011.
- Noah F Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565, 2022.
- Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.
- Ming Gui, Tianming Qiu, Fridolin Bauer, and Hao Shen. Laplace approximation for real-time uncertainty estimation in object detection. In

2022 *IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 409–415. IEEE, 2022.

Nicolas Guizard, Pierrick Coupé, Vladimir S Fonov, Jose V Manjón, Douglas L Arnold, and D Louis Collins. Rotation-invariant multi-contrast non-local means for ms lesion segmentation. *NeuroImage: Clinical*, 8: 376–389, 2015.

Axel Haase. Snapshot flash mri. applications to t1, t2, and chemical-shift imaging. *Magnetic resonance in medicine*, 13(1):77–89, 1990.

Joseph V Hajnal, David J Bryant, Larry Kasuboski, Pradip M Pattany, Beatrice De Coene, Paul D Lewis, Jacqueline M Pennock, Angela Oatridge, Ian R Young, and Graeme M Bydder. Use of fluid attenuated inversion recovery (flair) pulse sequences in mri of the brain. *Journal of computer assisted tomography*, 16(6):841–844, 1992a.

Joseph V Hajnal, Beatrice De Coene, Paul D Lewis, Christine J Baudouin, Frances M Cowan, Jacqueline M Pennock, Ian R Young, and Graeme M Bydder. High signal regions in normal white matter shown by heavily t2-weighted csf nulled ir sequences. *J Comput Assist Tomogr*, 16(4):506–513, 1992b.

Brian Hargreaves. Rapid gradient-echo imaging. *Journal of Magnetic Resonance Imaging*, 36(6):1300–1313, 2012.

K Harrison, H Pullen, C Welsh, O Oktay, J Alvarez-Valle, and R Jena. Machine learning for auto-segmentation in radiotherapy planning. *Clinical Oncology*, 34(2):74–88, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE*

transactions on pattern analysis and machine intelligence, 37(9):1904–1916, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016b.

Essam H Houssein, Marwa M Emam, Abdelmgeid A Ali, and Ponnuthurai Nagarathan Suganthan. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, 167:114161, 2021.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Shi Hu, Daniel Worrall, Stefan Knecht, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.

- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.
- Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.
- Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: A review. *BioMed Research International*, 2022, 2022.
- Nabil Ibtehaz and M Sohel Rahman. Multiresunet:rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting frame-

work for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Fabian Isensee, Paul F Jäger, Peter M Full, Philipp Vollmuth, and Klaus H Maier-Hein. nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*, pages 118–132. Springer, 2021.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2019.

- Andrew Jesson and Tal Arbel. Hierarchical mrf and random forest segmentation of ms lesions and healthy tissues in brain mri. *Proceedings of the 2015 longitudinal multiple sclerosis lesion segmentation challenge*, pages 1–2, 2015.
- Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.
- Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang, and Yan Xu. Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging*, 36(11):2376–2388, 2017.
- Amod Jog, Aaron Carass, Dzung L Pham, and Jerry L Prince. Multi-output decision trees for lesion segmentation in multiple sclerosis. In *Medical Imaging 2015: Image Processing*, volume 9413, pages 352–357. SPIE, 2015.
- Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.
- Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 48–56. Springer, 2019.
- Alain Jungo, Raphael Meier, Ekin Ermiş, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *International Conference on Medical Im-*

age Computing and Computer-Assisted Intervention, pages 682–690. Springer, 2018a.

Alain Jungo, Raphael Meier, Ekin Ermiş, Evelyn Herrmann, and Mauricio Reyes. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. *arXiv preprint arXiv:1806.03106*, 2018b.

Parvathy Jyothi and A Robert Singh. Deep learning models and traditional automated techniques for brain tumor segmentation in mri: a review. *Artificial intelligence review*, 56(4):2923–2969, 2023.

Alexandr A Kalinin, Vladimir I Iglovikov, Alexander Rakhlin, and Alexey A Shvets. Medical image segmentation using deep neural networks with pre-trained encoders. *Deep learning applications*, pages 39–52, 2020.

Konstantinos Kamnitsas, Liang Chen, Christian Ledig, Daniel Rueckert, Ben Glocker, et al. Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. *Ischemic stroke lesion segmentation*, 13:46, 2015.

Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2*, pages 138–149. Springer, 2016.

Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban

- remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–9, 2016.
- Reda Abdellah Kamraoui, Vinh-Thong Ta, Thomas Tourdias, Boris Mansencal, José V Manjon, and Pierrick Coupé. Deeplesionbrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis*, 76:102312, 2022.
- Guixia Kang, Beibei Hou, Yiyuan Ma, Fabrice Labeau, Zichen Su, et al. Acu-net: A 3d attention context u-net for multiple sclerosis lesion segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1384–1388. IEEE, 2020.
- Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.
- Jeremy Kawahara and Ghassan Hamarneh. Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In *International workshop on machine learning in medical imaging*, pages 164–171. Springer, 2016.
- Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brain-netcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.
- Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

- Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis*, 54:88–99, 2019.
- Hoel Kervadec, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *arXiv preprint arXiv:2004.06816*, 2020.
- Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.
- Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- Ashnil Kumar, Pradeeba Sridar, Ann Quinton, R Krishna Kumar, Dagan Feng, Ralph Nanan, and Jinman Kim. Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 791–794. IEEE, 2016.
- Viksit Kumar, Jeremy M Webb, Adriana Gregory, Max Denis, Duane D Meixner, Mahdi Bayat, Dana H Whaley, Mostafa Fatemi, and Azra Alizad. Automated and real-time segmentation of suspicious breast masses using convolutional neural network. *PloS one*, 13(5):e0195816, 2018.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised

- cell segmentation via self-generating pseudo-labels with consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 14–23. Springer, 2020.
- Sang Uk Lee, Seok Yoon Chung, and Rae Hong Park. A comparative performance study of several global thresholding techniques for segmentation. *Computer Vision, Graphics, and Image Processing*, 52(2): 171–190, 1990.
- Zihan Li, Yuan Zheng, Xiangde Luo, Dandan Shan, and Qingqi Hong. Scribblevc: Scribble-supervised medical image segmentation with vision-class embedding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3384–3393, 2023.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- Jasper Linmans, Stefan Elfwing, Jeroen van der Laak, and Geert Litjens. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83:102655, 2023.
- Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.

- Xiaoming Liu, Quan Yuan, Yaozong Gao, Kelei He, Shuo Wang, Xiao Tang, Jinshan Tang, and Dinggang Shen. Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. *Pattern recognition*, 122:108341, 2022.
- S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong Ki Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE transactions on medical imaging*, 14(4):711–718, 1995.
- Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.
- Wanxuan Lu, Dong Gong, Kun Fu, Xian Sun, Wenhui Diao, and Lingqiao Liu. Boundarymix: Generating pseudo-training images for improving segmentation with scribble annotations. *Pattern Recognition*, 117:107924, 2021.
- Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.
- Pietro Maggi, Martina Absinta, Matteo Grammatico, Luisa Vuolo, Giacomo Emmi, Giovanna Carlucci, Gregorio Spagni, Alessandro Barilaro, Anna Maria Repice, Lorenzo Emmi, et al. Central vein sign differentiates multiple sclerosis from central nervous system inflammatory vasculopathies. *Annals of neurology*, 83(2):283–294, 2018.
- Golnar K Mahani, Ruizhe Li, Nikolaos Evangelou, Stamatios Sotiropoulos, Paul S Morgan, Andrew P French, and Xin Chen. Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially con-

strained loss. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

Amirreza Mahbod, Chunliang Wang, and Orjan Smedby. Automatic multiple sclerosis lesion segmentation using hybrid artificial neural networks. *Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure-MICCAIMSSEG*, pages 29–36, 2016.

Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Rupert Ecker, and Isabella Ellinge. Skin lesion classification using hybrid deep neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1229–1233. IEEE, 2019.

Oskar Maier and Heinz Handels. Ms lesion segmentation in mri with random forests. *Proc. 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

Priyanka Malhotra, Sheifali Gupta, Deepika Koundal, Atef Zaguia, Weigayehu Enbeyle, et al. Deep neural networks for medical image segmentation. *Journal of Healthcare Engineering*, 2022, 2022.

Kanti V. Mardia and TJ Hainsworth. A spatial thresholding method for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 10(6):919–927, 1988.

Richard McKinley, Rik Wepfer, Tom Gundersen, Franca Wagner, Andrew Chan, Roland Wiest, and Mauricio Reyes. Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the*

Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2, pages 119–128. Springer, 2016.

Deval Mehta, Yaniv Gal, Adrian Bowling, Paul Bonnington, and Zongyuan Ge. Out-of-distribution detection for long-tailed and fine-grained skin lesion images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–742. Springer, 2022.

Afonso Menegola, Michel Fornaciali, Ramon Pires, Sandra Avila, and Eduardo Valle. Towards automated melanoma screening: Exploring transfer learning schemes. *arXiv preprint arXiv:1609.01228*, 2016.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4578–4585, 2019.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kertarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh. Learning to segment skin lesions from noisy annotations. In *Domain Adaptation and*

Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1, pages 207–215. Springer, 2019.

Niraj Mistry, Jennifer Dixon, Emma Tallantyre, Christopher Tench, Rasha Abdel-Fahim, Tim Jaspan, Paul S Morgan, Peter Morris, and Nikos Evangelou. Central veins in brain lesions visualized with high-field magnetic resonance imaging: a pathologically specific diagnostic biomarker for inflammatory demyelination in the brain. *JAMA neurology*, 70(5):623–628, 2013.

Niraj Mistry, Rasha Abdel-Fahim, Amal Samaraweera, Olivier Mouglin, Emma Tallantyre, Christopher Tench, Tim Jaspan, Peter Morris, Paul S Morgan, and Nikos Evangelou. Imaging central veins in brain lesions with 3-t t2*-weighted magnetic resonance imaging differentiates multiple sclerosis from microangiopathic brain lesions. *Multiple Sclerosis Journal*, 22(10):1289–1296, 2016.

Himanshu Mittal, Avinash Chandra Pandey, Mukesh Saraswat, Sumit Kumar, Raju Pal, and Garv Modwel. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications*, pages 1–26, 2021.

Mostafa Jabarouti Moghaddam and Hamid Soltanian-Zadeh. Medical image segmentation using artificial neural networks. *Artificial neural networks-Methodological advances and biomedical applications*, pages 121–138, 2011.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski,

- Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33:12756–12767, 2020.
- John P Mugler III and James R Brookeman. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3d mp rage). *Magnetic resonance in medicine*, 15(1):152–157, 1990.
- Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- John Muschelli, Elizabeth Sweeney, Jacob Maronge, and Ciprian Crainiceanu. Prediction of ms lesions using random forests. *Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure-MICCAI-MSSEG*, pages 45–50, 2016.
- Dong Nie, Han Zhang, Ehsan Adeli, Luyan Liu, and Dinggang Shen. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 212–220. Springer, 2016.
- Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- Tomoko Okuda, Yukunori Korogi, Yoshinori Shigematsu, Takeshi Sugahara, Toshinori Hirai, Ichiro Ikushima, Luxia Liang, and Mutsumasa

Takahashi. Brain lesions: when should fluid-attenuated inversion-recovery sequences be used in mr evaluation? *Radiology*, 212(3):793–798, 1999.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

Nikhil R Pal and Sankar K Pal. Object-background segmentation using new definitions of entropy. *IEE Proceedings E (Computers and Digital Techniques)*, 136(4):284–295, 1989.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

Jialin Peng and Ye Wang. Medical Image Segmentation With Limited Supervision: A Review of Deep Network Models. *IEEE Access*, 9: 36827–36851, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3062380. URL <https://ieeexplore.ieee.org/document/9363892/>.

R Prabha, M Razmah, S Sridevi, D Roopa, RM Asha, et al. A big wave of deep learning in medical imaging-analysis of theory and applications. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1321–1327. IEEE, 2022.

F Prados, MJ Cardoso, N Cawley, O Ciccarelli, CA Wheeler-Kingshott,

- and S Ourselin. Multi-contrast patchmatch algorithm for multiple sclerosis lesion detection. *ISBI-Longitudinal MS Lesion Segmentation Challenge*, pages 1–2, 2015.
- Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Gregory M Riedlinger, Subhajyoti De, and Dimitris N Metaxas. Weakly supervised deep nuclei segmentation using points annotation in histopathology images. In *International Conference on Medical Imaging with Deep Learning*, pages 390–400, 2019.
- Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2016.
- Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.
- Nicolas Robitaille, Simon Duchesne, et al. Label fusion strategy selection. *International journal of biomedical imaging*, 2012, 2012.
- Galina Rogova. Combining the results of several neural network classifiers. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 683–692, 2008.
- Torsten Rohlfing, Daniel B Russakoff, and Calvin R Maurer. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE transactions on medical imaging*, 23(8):983–994, 2004.

- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ” grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.
- Snehashis Roy, John A Butman, Daniel S Reich, Peter A Calabresi, and Dzung L Pham. Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172*, 2018.
- Kalpana Saini, ML Dewal, and Manojkumar Rohit. Ultrasound imaging and image segmentation in the area of ultrasound: a review. *International Journal of advanced science and technology*, 24, 2010.
- Natalia Salpea, Paraskevi Tzouveli, and Dimitrios Kollias. Medical image segmentation: A review of modern architectures. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022.
- Amal PR Samaraweera, Margareta A Clarke, Amy Whitehead, Yasser Falah, Ian D Driver, Robert A Dineen, Paul S Morgan, and Nikos Evangelou. The central vein sign in multiple sclerosis lesions is present irrespective of the t2* sequence at 3 t. *Journal of Neuroimaging*, 27(1):114–121, 2017.
- Pascal Sati, Jiwon Oh, R Todd Constable, Nikos Evangelou, Charles RG Guttmann, Roland G Henry, Eric C Klawiter, Caterina Mainero, Luca Massacesi, Henry McFarland, et al. The central vein sign and its clinical evaluation for the diagnosis of multiple sclerosis: a consen-

- sus statement from the north american imaging in multiple sclerosis cooperative. *Nature Reviews Neurology*, 12(12):714–722, 2016.
- Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.
- Suman Sedai, Bhavna Antony, Ravneet Rai, Katie Jones, Hiroshi Ishikawa, Joel Schuman, Wollstein Gadi, and Rahil Garnavi. Uncertainty guided semi-supervised segmentation of retinal layers in oct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 282–290. Springer, 2019.
- Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.
- Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, page 107441, 2023.
- Mehmet Sezgin and Bülent Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–166, 2004.
- Navid Shiee, Pierre-Louis Bazin, Arzu Ozturk, Daniel S Reich, Peter A Calabresi, and Dzung L Pham. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535, 2010.

- Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
- Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nasir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*, pages 10–18. Springer, 2016.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Arabella L Simpkin and Katrina A Armstrong. Communicating uncertainty: a narrative review and framework for future research. *Journal of general internal medicine*, 34:2586–2591, 2019.
- Bagher Sistaninejhad, Habib Rasi, Parisa Nayeri, et al. A review paper about deep learning for medical image analysis. *Computational and Mathematical Methods in Medicine*, 2023, 2023.
- John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.
- Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.
- Andrew J Solomon, Dennis N Bourdette, Anne H Cross, Angela Applebee, Philip M Skidd, Diantha B Howard, Rebecca I Spain, Michelle H Cameron, Edward Kim, Michele K Mass, et al. The contemporary

spectrum of multiple sclerosis misdiagnosis: a multicenter study. *Neurology*, 87(13):1393–1399, 2016.

Andrew J Solomon, Richard Watts, Daniel Ontaneda, Martina Absinta, Pascal Sati, and Daniel S Reich. Diagnostic performance of central vein sign for multiple sclerosis with a simplified three-lesion algorithm. *Multiple Sclerosis Journal*, 24(6):750–757, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Jean Stawiaski. A pretrained densenet encoder for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 105–115. Springer, 2018.

Martijn D Steenwijk, Petra JW Pouwels, Marita Daams, Jan Willem van Dalen, Matthan WA Caan, Edo Richard, Frederik Barkhof, and Hugo Vrenken. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (knn-ttps). *NeuroImage: Clinical*, 3:462–469, 2013.

Maddalena Strumia, Frank R Schmidt, Constantinos Anastasopoulos, Cristina Granziera, Gunnar Krueger, and Thomas Brox. White matter ms-lesion segmentation using a geometric brain model. *IEEE transactions on medical imaging*, 35(7):1636–1646, 2016.

Joshua Stuckner, Bryan Harder, and Timothy M Smith. Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset. *NPJ Computational Materials*, 8(1):200, 2022.

Nagesh Subbanna, Doina Precup, Douglas Arnold, and Tal Arbel. Image:

- iterative multilevel probabilistic graphical model for detection and segmentation of multiple sclerosis lesions in brain mri. In *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24*, pages 514–526. Springer, 2015.
- Farhana Sultana, Abu Sufian, and Paramartha Dutta. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowledge-Based Systems*, 201:106062, 2020.
- Jesse Sun, Fatemeh Darbehani, Mark Zaidi, and Bo Wang. Saunet: Shape attentive u-net for interpretable medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 797–806. Springer, 2020.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions (2014). *arXiv preprint arXiv:1409.4842*, 10, 2014.
- Vinh-Thong Ta, Rémi Giraud, D Louis Collins, and Pierrick Coupé. Optimized patchmatch for near real time and accurate label fusion. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14–18, 2014, Proceedings, Part III 17*, pages 105–112. Springer, 2014.
- Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

- EC Tallantyre, JE Dixon, I Donaldson, T Owens, PS Morgan, PG Morris, and N Evangelou. Ultra-high-field imaging distinguishes ms lesions from asymptomatic white matter lesions. *Neurology*, 76(6):534–539, 2011.
- Emma C Tallantyre, Paul S Morgan, Jennifer E Dixon, Ali Al-Radaideh, Matthew J Brookes, Nikos Evangelou, and Peter G Morris. A comparison of 3t and 7t in the detection of small parenchymal veins within ms lesions. *Investigative radiology*, 44(9):491–494, 2009.
- Lucas O Teixeira, Rodolfo M Pereira, Diego Bertolini, Luiz S Oliveira, Loris Nanni, George DC Cavalcanti, and Yandre MG Costa. Impact of lung segmentation on the diagnosis and explanation of covid-19 in chest x-ray images. *Sensors*, 21(21):7116, 2021.
- Atsushi Teramoto, Hiroshi Fujita, Osamu Yamamuro, and Tsuneo Tamaki. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical physics*, 43(6Part1):2821–2827, 2016.
- Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. *arXiv preprint arXiv:1802.06455*, 2018.
- Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.
- Michael J Trimpl, Sergey Primakov, Philippe Lambin, Eleanor PJ Stride, Katherine A Vallis, and Mark J Gooding. Beyond automatic medical image segmentation—the spectrum between fully manual and fully

automatic delineation. *Physics in Medicine & Biology*, 67(12):12TR01, 2022.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villa, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. Multiple sclerosis lesion detection and segmentation using a convolutional neural network of 3d patches. *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 75:84–88, 2016.

Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

Sergi Valverde, Mostafa Salem, Mariano Cabezas, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21:101638, 2019.

Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

- F Vera-Olmos, H Melero, and N Malpica. Random forest for multiple sclerosis lesion segmentation. *Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure-MICCAI-MSSEG*, pages 81–86, 2016.
- Juan Wang and Bin Xia. Bounding box tightness prior for weakly supervised image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 526–536. Springer, 2021.
- Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 2022.
- Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.
- Guorong Wu, Minjeong Kim, Qian Wang, Yaozong Gao, Shu Liao, and Dinggang Shen. Unsupervised deep feature learning for deformable registration of mr brain images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 649–656. Springer, 2013.
- Kai Wu, Bowen Du, Man Luo, Hongkai Wen, Yiran Shen, and Jianfeng Feng. Weakly supervised brain lesion segmentation via attentional representation learning. In *Medical Image Computing and Computer*

Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22, pages 211–219. Springer, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Chenyang Xu, Dzung L Pham, and Jerry L Prince. Image segmentation using deformable models. *Handbook of medical imaging*, 2(20):0, 2000.

Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, I Eric, and Chao Chang. Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1626–1630. IEEE, 2014.

Guanyu Yang, Chuanxia Wang, Jian Yang, Yang Chen, Lijun Tang, Pengfei Shao, Jean-Louis Dillenseger, Huazhong Shu, and Limin Luo. Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal cta images. *BMC medical imaging*, 20:1–12, 2020.

Lin Yang, Yizhe Zhang, Zhuo Zhao, Hao Zheng, Peixian Liang, Michael TC Ying, Anil T Ahuja, and Danny Z Chen. Boxnet: Deep learning based biomedical image segmentation using boxes only annotation. *arXiv preprint arXiv:1806.00593*, 2018.

Xiao Yang, Roland Kwitt, and Marc Niethammer. Fast predictive image registration. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MIC-*

CAI 2016, Athens, Greece, October 21, 2016, Proceedings 1, pages 48–57. Springer, 2016.

Moi Hoon Yap, Gerard Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017.

Jong Chul Ye and Woon Kyoung Sung. Understanding geometry of encoder-decoder cnns. In *International Conference on Machine Learning*, pages 7064–7073. PMLR, 2019.

Michael Yeung, Leonardo Rundo, Yang Nan, Evis Sala, Carola-Bibiane Schönlieb, and Guang Yang. Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. *Journal of Digital Imaging*, 36(2):739–752, 2023.

Xiao-Xia Yin, Le Sun, Yuhan Fu, Ruiliang Lu, Yanchun Zhang, et al. U-net-based medical image segmentation. *Journal of Healthcare Engineering*, 2022, 2022.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

Zimin Yu, Li Yu, Weihua Zheng, and Shunfang Wang. Eiu-net: Enhanced feature extraction and improved skip connections in u-net for skin lesion segmentation. *Computers in Biology and Medicine*, 162:107081, 2023.

- Muhammad Zaid, Shafaqat Ali, Mohsen Ali, Sarfaraz Hussein, Asma Saadia, and Waqas Sultani. Identifying out of distribution samples for skin cancer and malaria images. *Biomedical Signal Processing and Control*, 78:103882, 2022.
- Hang Zhang, Jinwei Zhang, Chao Li, Elizabeth M Sweeney, Pascal Spincemaille, Thanh D Nguyen, Susan A Gauthier, Yi Wang, and Melanie Marcille. All-net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation. *NeuroImage: Clinical*, 32:102854, 2021.
- Huahong Zhang, Alessandra M Valcarcel, Rohit Bakshi, Renxin Chu, Francesca Bagnato, Russell T Shinohara, Kilian Hett, and Ipek Oguz. Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 338–346. Springer, 2019a.
- Ning Zhang, Susan Francis, Rayaz A Malik, and Xin Chen. A spatially constrained deep convolutional neural network for nerve fiber segmentation in corneal confocal microscopic images using inaccurate annotations. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 456–460. IEEE, 2020.
- Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.
- Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by

deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15 (5):749–753, 2018.

Zhenxi Zhang, Jie Li, Zhusi Zhong, Zhicheng Jiao, and Xinbo Gao. A sparse annotation strategy based on attention-guided active learning for 3d medical image segmentation. *arXiv preprint arXiv:1906.07367*, 2019b.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109 (5):820–838, 2021.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

Appendices

Appendix A

Uncertainty Map

Interpretations in Medical Image Segmentation

Tables A.1, A.2, and A.3 show the non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in ISIC2018, BraTS and BUS17 datasets, respectively.

Comparison	H	P-value
Baseline vs. Aleatoric Loss	0	0.4962
Baseline vs. UCF Loss	0	0.7639
Baseline vs. ELF Loss	0	0.7008
Baseline vs. UCF + AL	0	0.0804
Baseline vs. ELF + AL	0	0.8104
Aleatoric Loss vs. UCF Loss	0	0.4925
Aleatoric Loss vs. ELF Loss	0	0.5496
Aleatoric Loss vs. UCF + AL	1	0.0012
Aleatoric Loss vs. ELF + AL	0	0.3909
UCF Loss vs. ELF Loss	0	0.9321
UCF Loss vs. UCF + AL	0	0.2313
UCF Loss vs. ELF + AL	0	0.6032
ELF Loss vs. UCF + AL	1	0.0030
ELF Loss vs. ELF + AL	0	0.3591
UCF + AL vs. ELF + AL	0	0.0607

Table A.1: Non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in ISIC2018 dataset. Bold value indicates that the result of Wilcoxon test is significant (significance rate is 0.01).

Comparison	H	P-value
Baseline vs. Aleatoric Loss	0	0.4649
Baseline vs. UCF Loss	0	0.0866
Baseline vs. ELF Loss	1	0.0082
Baseline vs. UCF + AL	0	0.9051
Baseline vs. ELF + AL	0	0.2317
Aleatoric Loss vs. UCF Loss	1	0.0358
Aleatoric Loss vs. ELF Loss	1	0.0388
Aleatoric Loss vs. UCF + AL	0	0.2105
Aleatoric Loss vs. ELF + AL	0	0.3675
UCF Loss vs. ELF Loss	0	0.2594
UCF Loss vs. UCF + AL	0	0.0511
UCF Loss vs. ELF + AL	0	0.3290
ELF Loss vs. UCF + AL	1	0.0146
ELF Loss vs. ELF + AL	0	0.1740
UCF + AL vs. ELF + AL	0	0.1569

Table A.2: Non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in BraTS dataset. Bold value indicates that the result of Wilcoxon test is significant (significance rate is 0.01).

Comparison	H	P-value
Baseline vs. Aleatoric Loss	0	0.8107
Baseline vs. UCF Loss	0	0.9039
Baseline vs. ELF Loss	0	0.9133
Baseline vs. UCF + AL	0	0.8313
Baseline vs. ELF + AL	0	0.9058
Aleatoric Loss vs. UCF Loss	0	0.8405
Aleatoric Loss vs. ELF Loss	0	0.9811
Aleatoric Loss vs. UCF + AL	0	0.7771
Aleatoric Loss vs. ELF + AL	0	0.9359
UCF Loss vs. ELF Loss	0	0.8107
UCF Loss vs. UCF + AL	0	0.9811
UCF Loss vs. ELF + AL	0	0.7475
ELF Loss vs. UCF + AL	0	0.9133
ELF Loss vs. ELF + AL	0	0.8684
UCF + AL vs. ELF + AL	0	0.8107

Table A.3: Non-parametric Wilcoxon signed rank significance test based on the Dice Coefficient performance measure for test subjects in BUS17 dataset. Bold value indicates that the result of Wilcoxon test is significant (significance rate is 0.01).

Appendix B

Understanding Multiple Sclerosis: Challenges and Deep Convolutional Neural Network MS Lesion Segmentation

B.1 Model Training and Parameter Settings

The baseline architecture of the MUnet reference model is Unet with 5 layers in both the encoder and decoder paths. The models are trained for 50 epochs with Adam optimiser and learning rate of 0.0001. To calculate the uncertainty map during the training and at inference time, the dropout rate and the number of stochastic forward passes were set to $d = 0.05$ and $M = 20$. The input images for segmentation using MUnet were downsized to (140 , 140 , 170) due to computational reasons and the model was only trained on 2D slices of the MRI images.

The DeepMedic [Kamnitsas et al., 2016] model contains 3 resolution pathways, one with the original volumes and two parallel pathways with down-sampling scales of (3, 3, 3) and (5, 5, 5). Each pathway contains 8 layers. Two Fully Connected (FC) hidden layers are added before the final classification layer. The dropout rates for the pathways, the last FC layer and the classification layer are 0, 0.5 and 0.5, respectively, and the size of the training segments are (37, 37, 37). The models are trained for 700 epochs with Adam optimiser and 0.001 learning rate and cross entropy loss function.

The parameter settings and corresponding loss function for each loss function are listed in Table B.1. All the parameters are chosen experimentally and fixed throughout the experiments.

Model	LossFunction	Parameter settings
MUnet (baseline)	$\mathcal{L}_{ce}$	-
MUnet+UCF	$\mathcal{L}_{UCF}$	$\alpha = 0.1, \beta = 3$
MUnet+ELF	$\mathcal{L}_{ELF}$	$\alpha = 0.2, \beta = 3$
MUnet+UCF+AL	$\mathcal{L}_{UCF+AL}$	$\alpha = 0.1, \beta = 3, \lambda = 1, sampling = 10$
MUnet+ELF+AL	$\mathcal{L}_{ELF+AL}$	$\alpha = 0.2, \beta = 3, \lambda = 0.1, sampling = 10$

Table B.1: Model Configurations with their Corresponding Loss Functions and Hyper-parameter Settings For MS Dataset